2018 IEEE International Conference on Robotics and Automation (ICRA)
May 21-25, 2018, Brisbane, Australia

Beyond Pixels: Leveraging Geometry and Shape Cues for Online Multi-Object Tracking

Sarthak Sharma[1*], Junaid Ahmed Ansari[1*], J. Krishna Murthy[1] and K. Madhava Krishna[1]

Abstract— **This paper introduces geometry and object shape and pose costs for multi-object tracking in urban driving scenarios. Using images from a monocular camera alone, we devise pairwise costs for object tracks, based on several 3D cues such as object pose, shape, and motion. The proposed costs are agnostic to the data association method and can be incorporated into any optimization framework to output the pairwise data associations. These costs are easy to implement, can be computed in real-time, and complement each other to account for possible errors in a tracking-by-detection framework. We perform an extensive analysis of the designed costs and empirically demonstrate consistent improvement over the state-of-the-art under varying conditions that employ a range of object detectors, exhibit a variety in camera and object motions, and, more importantly, are not reliant on the choice of the association framework. We also show that, by using the simplest of associations frameworks (two-frame Hungarian assignment), we surpass the state-of-the-art in multi-object-tracking on road scenes. More qualitative and quantitative results can be found at `https://junaidcs032.github.io/Geometry_ObjectShape_MOT/`.**

I. INTRODUCTION

Object tracking in road scenes is an important component of urban scene understanding. With the advent and subsequent surge in autonomous driving technologies, accurate multi-object trackers are desirable in several tasks such as navigation and planning, localization, and traffic behavior analysis.

In this paper, we focus on designing a simple and fast, yet accurate and robust solution to the Multi-Object Tracking (MOT) problem in an urban road scenario. The dominant approach to multi-object tracking is tracking-by-detection, where the entire process is divided into two phases. The first phase comprises object detection, where bounding-boxes of objects of interests are obtained in each frame of the video sequence. The second phase is the data association phase, which is often the hardest step in the tracking-by-detection paradigm. Several factors such as spurious or missing detections, repeat detections, or occlusions and target interactions are confounding factors in this data association phase.

Although several approaches [1], [2], [3], [4], [5] exist for accurate online tracking of moving vehicles from a moving camera, most of them [6], [2] use handcrafted cost functions that are either based on primitive features such as bounding box position in the image and color histograms, or are highly sophisticated and non-intuitive in design (eg. ALFD [6]). On the other hand, we propose costs that are intuitive, easy to compute and implement, and provide complementary cues about the target.

We exploit the fact that road scenes have a unique geometry and use this prior information to design costs. The proposed costs capture 3D cues arising from this scene geometry, as well as appearance based information. Further, we introduce a novel cost that captures similarity of 3D shapes and poses of target hypotheses. To this end we leverage recent work on shape-priors for object detection and localization from monocular sequences [7], [8]. To the best of our knowledge, such pairwise costs have not been incorporated in multi-object tracking frameworks.

The efficacy of the monocular 3D cues is best portrayed in Fig.1. In this figure the first two rows illustrate the objects with their bounding boxes in two successive frames at t and $t+1$. Upon lifting the objects at t to 3D and ballooning their locations to account for large uncertainties in ego motion, we project them into the image observed at $t + 1$. This gated/overlapping area shown in their respective colors in the last row of Fig.1 reduces the search area for each such object significantly thereby reducing the pairwise costs. By backprojecting detections that lie only within this gated area into 3D and ascertaining data association costs based on 3D volume overlaps significantly improves tracking accuracy even with a straight forward Hungarian data association scheme.

The proposed costs are not too dependent on the choice of data association framework. We demonstrate the superiority of the proposed costs over monocular video sequences of urban road scenes that capture a wide range of camera and target motions, and also consistent improvement over other costs regardless of the choice of the object detector. We perform an extensive evaluation of various modes of the proposed costs on the KITTI Tracking benchmark [9] and obtain state-of-the-art performance, beating previous approaches by using a simple two-frame Hungarian association scheme. The approach is tested on KITTI online evaluation sever and outperforms the previous published approaches significantly. Naturally, more complex data association schemes, such as network flow based algorithms [10], [11], [12], [13] can result in much better performance boosts upon incorporation of the proposed pairwise costs.

The paper contributes as follows.

1) It introduces novel data association cues based on single view reconstruction of objects that results in best tracking performance reported thus far in KITTI

*The first two authors contributed equally to this work.

The authors are with the Robotics Research Center, KCIS, IIIT Hyderabad, India. `junaid.ansari@research.iiit.ac.in` This work was supported by grants made available by Qualcomm Innovation Fellowship India, 2017.

978-1-5386-3081-5/18/$31.00 ©2018 IEEE

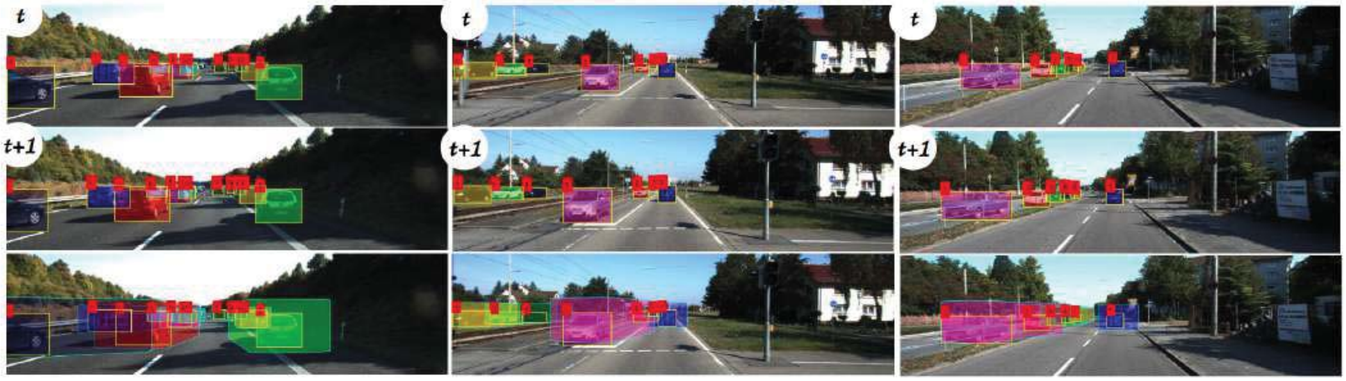

Fig. 1. An illustration of the proposed method. The first two rows show objects tracks in frames t and $t+1$. The bottom row depicts how 3D position and orientation information is propagated from frame t to frame $t+1$. This information is used to specify search areas for each object in the subsequent frame, and this greatly reduces the number of pairwise costs that are to be computed.

training datasets. It outperforms the nearest reported values in training data [14], [1], [6], by at-least 12% . The approach is tested on the KITTI Tracking online evaluation server where it outperforms the published approaches by a margin of over 6%.

2) It showcases that such improvements are sufficiently detector agnostic and repeatable over baseline appearance tracking based on object detectors such as [15], [16] through ablation studies
3) Finally it also identifies a role for 3D pose and shape cues where they play a role in improving tracking performance.

Monocular 3D cues especially based on single view geometry can often be unreliable. However when computed effectively they can be used reliably and repeatably even in challenging sequences such as KITTI. This constitutes the central theme of this effort.

II. Related Work

In this section, we review relevant work on multi-object tracking, and compare and contrast it with the proposed approach.

A. Global Tracking

Many approaches to tackle the association problem are *global* [10], [12], [17], [11], [18], [19], in the sense that they assume detections from all frames are available for processing. Most global methods operate by mapping the tracking problem to a min-cost network flow problem. The original idea was proposed in [10] and also provides for a method for explicit occlusion reasoning. An efficient variant is an approach based on generalized minimum clique graphs [12], where associations are solved for one object at a time while other objects are implicitly incorporated. Another section of global methods attempts to construct small chunks of trajectories (called tracklets), and compose them hierarchically to form longer trajectories, rather than solving for a min-cost flow over a densely connected graph.

B. Online MOT

In contrast to this, online trackers [4], [20], [3], [21] do not assume any knowledge of future frames and operate greedily, only with the data available upto the current instant. Such trackers often formulate the association problem as that of bipartite matching, and solve it via the Hungarian algorithm. A recent variant proposes near-online trackers [6], in an attempt to provide the best of both worlds, i.e., to combine the capability of global methods to handle long-term occlusions and still achieve very low output latencies. Gieger et al [13] propose a memory and computation cost bound variant of network flow using dynamic programming.

Both these paradigms rely on handcrafted pairwise costs being fed into the association framework. Most of these are sophisticated in design and do not end up capturing 3D information that is easily available in road scenes.

C. Learning Costs for MOT

Significant attention has also been devoted to the task of learning pairwise costs for target tracking problems. In [3], a structured SVM was used to learn pairwise costs for a bipartite matching data association framework. Other works have used graphical models, divide and conquer strategies and also learn unary costs. A more recent work [1] learns all costs using a deep neural network. On the other hand, we show that our simple, yet clean and efficient cost function designs significantly improve performance without the need of extensive hyperparameter search or cost learning.

III. Problem Formulation

We adopt the tracking-by-detection paradigm where we assume that we are provided with a monocular video sequence of F frames $\{I_f\}$ for $f \in \{1..F\}$, and a set of object detections D_f for each frame I_f. Each detection set consists of object detections $\{D_f^i\}$, where $i \in \{1..N_f\}$ (N_f is the number of detections in frame f). Note that D_f can also be an empty set, in the case where no objects are detected in a frame. Each detection D_f^i is parametrized as $D_f^i = (x_f^i, y_f^i, w_f^i, h_f^i, s_f^i)$, where (x_f^i, y_f^i) corresponds to the top-left corner of the detection box in the image, w_f^i

is the bounding box width, h_f^i is the bounding box height, and s_f^i is the detectors confidence in the bounding box (greater value indicates higher confidence). The multi-object tracking problem is to associate each bounding box to a target trajectory T_k such that the following constraints are met.

- Each target trajectory T_k comprises of a set of bounding boxes (all from different frames) belonging to a unique target in the scene.
- There are exactly as many trajectories K as there are targets to be tracked.
- In all frames where a target is visible, it is detected and assigned to the corresponding unique trajectory for the object.
- All spurious bounding box detections are unassigned to any target trajectory.

The tracking problem formulated above is usually solved in a min-cost network flow framework (global tracking), a moving window dynamic programming framework (near-online tracking) or a bipartite matching framework (online tracking). Note that these are not the only available frameworks, but a representative set of most tracking approaches. All these frameworks (and the others not mentioned here) use pairwise costs to define affinity across pairs of detections. The association framework then computes a Maximum A Posteriori (MAP) estimate of the target trajectory T_k, given the detection hypotheses $D = (D_1, D_2, ...D_f)$ and an affinity matrix that gives the likelihood of each detection in each frame corresponding to every detection in every other frame.

IV. Geometry and Object Shape Costs

The core contribution of this paper is to design intuitive pairwise costs that are efficient to compute, and are accurate for tracking. We focus on urban driving scenarios and demonstrate how the geometry of urban road scenes can be exploited to infer 3D cues for tracking.

Typical costs in tracking algorithms include bounding box locations, trajectory priors, optical flow, bounding box overlap, and appearance information (color histograms or path-based cross-correlation measures). These costs require careful handcrafting, finetuning, and hyperparameter estimation. We propose to use a set of simple complementary costs that are readily available from recent monocular 3D object localization systems [7], [8]. We also introduce a novel cost based on the 3D shape and pose of the target. We show that this cost, apart from improving data association performance, also assists in discarding false detections without incurring large computational overhead.

A. System Setup

We focus on autonomous driving scenarios, where the video sequence is from a monocular camera mounted on a car moving on the road plane, and the targets to be tracked are also moving on the road. Feature based odometry is run on a background thread (for rough frame-to-frame motion estimation). Also, we make use of a recent approach that goes beyond bounding boxes and estimates the 3D shape and pose of objects, given just a single image [7]. This is done by lifting discriminative parts in 2D (*keypoints*) to 3D. These keypoints are a set of points chosen so that they are common across all object instances (eg. for a car, we have centers of wheels, headlights, taillights, etc). The authors use a CNN architecture [8] to localize these keypoints in 2D, given a detection.

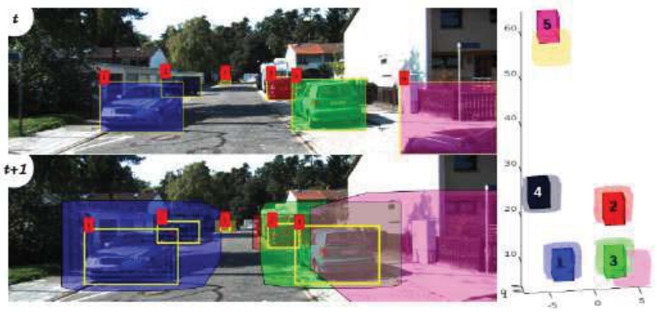

Fig. 2. Illustrating the concept of 3D-2D and 3D-3D costs Two subsequent frames, t and t+1, are shown in the left. For each detection in the frame t, we compute and propagate (with uncertainty) its 3D bounding box in the next frame t + 1. These boxes are projected to 2D in the frame t+1. The intersection between the detection boxes of t+1 and these projections constitutes the 3D-2D cost. The intersection of the 3D bounding boxes in 3D constitute the 3D-3D cost as shown in the right; propagated bounding boxes are colored with their respective 2D box in frame t and 3D bounding boxes of detections in frame t+1 are numbered respectively.

The 3D shape of the object is parametrized as the sum of the mean shape (for the object category) and a linear combination of so-called *basis shapes*. Mathematically,

$$S = \bar{S} + \sum_{b=1}^{b} \lambda_b V_b \tag{1}$$

where S is the shape of a particular instance, $\bar{S}$ is the mean shape for the object category, and V_b is the deformation basis (a set of eigenvectors) that characterizes deformation directions of the mean shape. We use the same model in [7] and denote the shape vector of an object by $\Lambda = [\lambda_1..\lambda_B]^T$, where B is the number of vectors in the deformation basis (typically, $B = 5$).

The pipeline in [7] also estimates the 3D pose of the object, which is parametrized as an axis-angle vector ω. Moreover, an estimate of object dimensions (height, width, and length) is also returned.

B. 3D-2D Cost

Given the height h_{cam} of the camera above the ground, assuming that the bottom line of each bounding box detection d_f^i in frame f is on the road plane, a depth estimate of the car in the current camera coordinates can be obtained by back projection via the road plane as in [22], using

$$X_f = \pi_G^{-1}(x) - \frac{h\mathbf{K}^{-1}x}{n^T\mathbf{K}^{-1}x} \tag{2}$$

where x is the bottom center of the detected bounding box, $\mathbf{K}$ is the camera intrinsic matrix and π_G^{-1} is used as shorthand for backprojection via the ground plane. This backprojection equation is only accurate when x is known precisely, which is not usually the case. Hence, we estimate the uncertainty in 3D location of X_f by using a linearized version of (2) and assuming that the detector confidence is an isotropic 2D Gaussian, i.e., $(x^i_f, y^i_f)^T \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{2\times 2})$. This region is expanded (anisotropically) by the estimates of the target dimensions returned by the system [7].

Now, assume we have another detection $d^j_{f'}$ in frame f' with which we wish to compute the pairwise affinity of d^i_f. We obtain a rough estimate of the camera motion from frame f to frame f' using a feature-based odometry thread running in the background. Using this estimate of the camera motion, we transport X_f to the camera coordinates of frame f', while duly accounting for the uncertainty in camera motion estimate, and in the backprojection via the road plane. The obtained coordinates $X_{f'}$ are then projected down to the image frame f' to obtain a 2D search area in which potential matches for X_f are expected to be found, as shown in the frame $t+1$ of Fig.2. Mathematically, the 3D-2D cost for two detections d^i_f and $d^j_{f'}$ is defined as follows

$$\mathcal{C}_{3D2D}(d^i_f, d^j_{f'}) = 1 - \frac{(\pi(g(\xi, \phi(\pi_G^{-1}(y^i_f), s^i_f)) \cap b^j_{f'}))}{b^j_{f'}} \quad (3)$$

Intuitively, this cost measures a (weighted) overlap of the 2D region in which the target is expected in frame f' and the detection $d^j_{f'}$. π denotes the projection operator that projects a 3D point to image pixel coordinates. $g(\xi, X)$ denotes a rigid-body motion $\xi \in se(3)$ applied to a 3D point $X \in \mathbb{R}^3$. $\phi(X, s)$ denotes the function that estimates the uncertainty of the 3D point S according to a linearized form of (2) and the detector confidence s.

Most importantly, this cost is evaluated only for detections $d^j_{f'}$ that lie inside the expected target area $(\pi(g(\xi, \phi(\pi_G^{-1}(y^i_f), s^i_f))$. This significantly reduces the number of comparisons needed to be made among target pairs.

C. 3D-3D cost

Although useful in reducing the number of candidate detections to be evaluated, the 3D-2D cost has frequent confounding cases. This is because, we still measure overlap in the image space. To mitigate this drawback, we define a 3D-3D cost, which, instead of measuring 2D overlap, measures overlap in 3D,as shown in Fig.2 (right side). Here, we backproject each candidate $d^j_{f'}$ via the road plane, and measure overlap with respect to the transformed 3D volume from frame f given by $g(\xi, \phi(\pi_G^{-1}(y^i_f), s^i_f)$. The 3D-3D cost for two detections d^i_f and $d^j_{f'}$ is defined as

$$\mathcal{C}_{3D3D}(d^i_f, d^j_{f'}) = 1 - \frac{g(\xi, \phi(\pi_G^{-1}(y^i_f), s^i_f))}{\phi(\pi_G^{-1}(b^j_{f'}), s^j_{f'})} \quad (4)$$

In order to speed up evaluation of 3D overlap, we exploit the inherent geometry of road scenes. Since all objects of interest are on the road plane (the XZ plane in our case), it is sufficient to measure overlap in the XZ plane. This is because all objects are at nearly constant heights above the ground and hence have similar overlap in the Y direction.

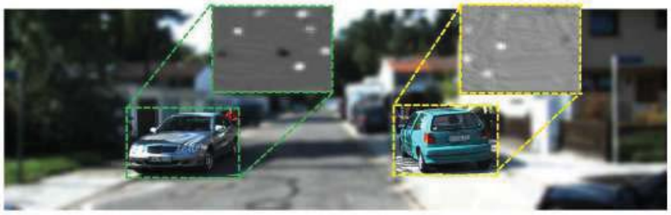

Fig. 3. Weighted combination of the features captured by the hourglass network. Such a descriptor is able to capture the dissimilarity between different detections.

D. Appearance Cost

In [8], the authors train a stacked-hourglass CNN architecture to localize a discriminative set of keypoints on an image. This deep CNN architecture captures various discriminative features for each detection, along with the keypoint evidence. We use weighted combination of activation maps from the output of the layers of the hourglass network as a feature descriptor for each detection, as shown in Fig.3 and compute a similarity score between detections using the L2 Norm between descriptors from the image patch inside each of the bounding boxes. If $\psi(.)$ denotes the feature descriptor of each detection, the appearance cost is defined as

$$\mathcal{C}_{app}(d^i_f, d^j_{f'}) = \eta_{app} \|\psi(d^i_f) - \psi(d^j_{f'})\|_2^2 \quad (5)$$

where η_{app} is a normalization constant.

E. Shape and Pose Cost

We use a novel shape and pose cost based on the single image shape and pose returned by the pipeline of [7] . Shape is parameterized as a vector comprising of deformation coefficients $\Lambda = [\lambda_1..\lambda_B]^T$, where B is the number of deformation basis vectors (usually 5). Each possible value of Λ denotes a unique class of object instances and hence carries useful information about the 3D shape of the target. For instance varying certain parameters of Λ may represent a shape that is more SUV-like than Sedan-like, and so on. Pose is parametrized as an axis-angle vector ω. For detections d^i_f and $d^j_{f'}$, the shape and pose cost is specified as

$$\mathcal{C}_s(d^i_f, d^j_{f'}) = \eta_s \|\Lambda(d^i_f) - \Lambda(d^j_{f'})\|_2^2 + \eta_p \|\omega(d^i_f) - \omega(d^j_{f'})\|_2^2 \quad (6)$$

where η_s and η_p are normalization constants.

The overall pairwise cost term is a weighted linear combination of all the aforementioned cost. The weights of the linear combination are determined by four-fold cross validation on the train set.

V. Results

In this section, we present an account of the experiments we performed, and we report and analyze the findings thereof. In nutshell, we evaluate our tracking framework on a variety of challenging urban driving sequences and demonstrate a substantial performance boost over the state-of-the-art in multi-object tracking, by using the simplest of tracking frameworks, viz. bipartite matching using the Hungarian algorithm.

A. Dataset

We evaluate the proposed multi-object tracking framework on the popular KITTI Tracking benchmark on both training as well as testing dataset. [9]. As prescribed in [1], [6], [9], we divide the training dataset, which contains 21 sequences, into four splits,for cross validation. The splits are chosen so that each split contains a similar distribution of number of vehicles per sequence, occlusion and truncation levels, and relative motion patterns between the camera and the target. The cross validation helps us to tune the weight for each of the proposed costs to compute the final cost matrix. The best performing combination of these weighted costs are used for reporting the result on the KITTI Tracking benchmark. Multiple vehicles moving with varying speeds, variance in the ego camera motion, and target objects appearing in non conforming locations in frames make the KITTI Tracking dataset [9] a truly challenging one. We report results on the *Car* class.

B. Evaluation Metrics

To evaluate the performance of our approach, we adopt the widely used CLEAR MOT metrics [25]. The overall performance of the tracker is summed up in two intuitive metrics, viz. Multi-Object Tracking Accuracy (MOTA) and Multi-Object Tracking Precision (MOTP). While MOTA is concerned with tracking accuracy, MOTP deals with object localization precision.

C. System Overview

The proposed approach is a tracking-by-detection approach and hence assumes per-frame bounding box detections as input. We choose two recent object detectors — Recurrent Rolling Convolution (RRC) [15] and SubCNN [16]. Each of these detectors provides multiple detections per frame. A threshold is applied on the detection scores and those detections whose confidence scores are lower than the threshold are pruned. In addition to this, we run a non-maxima suppression (NMS) scheme to subdue multiple detections around the same object. These detections are used to compute pairwise costs as outlined in the previous section. These pairwise costs constitute a cost matrix that is used for a bipartite matching algorithm that associates detections across two frames. In practice, bipartite matching is performed using the $O(n^3)$ Hungarian algorithm [26].

D. Approaches Considered

We evaluate the proposed framework against the current best competitors on the KITTI Tracking dataset. We consider approaches [1], [6], [21], [13].

E. Performance Evaluation

We evaluate the performance of our approach on the current best competitors on the KITTI Tracking Benchmark. While [6], [21], [13] rely on complex handcrafted costs, [1] learns all unary and pairwise costs that are input to a network flow based tracker. Moreover, the data association steps of [6], [21], [13] rely on complex optimization routines. The proposed approach is also evaluated on the KITTI Tracking evaluation sever.

Table I,where we compare our two-frame based approach with the other competitors using the best performing object detector in the form of [15] and a judicious combination of such appearance, 3D, pose and shape cues best possible results on KITTI training sequence are achieved in terms of MOTA (91.4%) and MOTP (89.84%). Although our method suffers from ID switches and fragmentations, this is typical of online trackers; more so of two-frame greedy trackers. Using the proposed pairwise costs in a slightly more sophisticated tracker such as [6], [13] will naturally reduce ID switches and fragmentations also.

Table II,where we compare our two-frame based approach with the other published approaches on the KITTI Tracking online server. We outperform the next best competitor by a margin of (6%) on the test set, achieving state of the art results in the form of MOTA (84.24%), MOTP (85.73%), MT (73.23%) and ML (2.77%).

F. Ablation Study

We then perform a thorough ablation analysis of various cues used for computing pairwise costs across two distinct object detectors: RRC [15] and SubCNN [16]. Results are summarized in Table III. This analysis captures the importance of each of the proposed cue and demonstrates that the combination of all these is crucial for overall performance. Notice how each cue improves the performance of our system in terms of MOTA ,ID switches and fragmentations. Even with underperforming detectors such as [16], there is a tangible performance boost by using a combination of monocular 3D cues. This is portrayed in ablation analysis of SubCNN detectors in Table III. Furthermore the repeatability of performance gain using these novel cues over any baseline detection methods is also delineated.

There exist subsequences where the role played by shape and pose cues become relevant. While in a typical road scene involving lane driving the pose cues are not discriminatory (as the vehicles are aligned with the lane direction), they become discerning enough in areas such as intersections, round abouts where pose and viewpoint changes are heterogeneous. This is showcase in Table IV. Here, we select particular frames from the KITTI Tracking dataset, which have images containing cars moving at intersections, which captures different viewpoints and shapes of cars. Using detections from a weak detector [16] and a simplistic combination 2D-2D cues along with shape and pose cue of the car performs better than the stand alone 2D cue, for sequences which have cars with various viewpoints over the frames.

G. Qualitative Results

Finally, we present qualitative results from challenging sequences in Fig.4 and Fig.5. These results clearly indicate the ability of the proposed pairwise costs to disambiguate and track across viewpoint variations, clutter, and varying relative motion between the camera and the target.

TABLE I

Results on the KITTI Tracking train set. Tracking Accuracy(MOTA) and Precision(MOTP), Mostly Tracked(MT),Partly Tracked(PT), Mostly Lost(ML) , True Positives(TP) , False Positives(FP), ID-switches(IDS), Fragmentation(FRAG)

	MOTA	MOTP	Recall	Precision	MT	PT	ML	TP	FP	IDS	FRAG
CIWT [21]	74.38	82.85	-	-	49.59	40.68	9.80	-	-	**26**	**131**
Ours (On split of [21])	**91.75**	**89.90**	94.83	98.62	**88.61**	10.39	**0.9**	9814	137	93	151
Deep Network Flow [1]	74.11	-	84.74	92.05	61.73	-	-	-	-	**29**	**335**
NOMT [6]	73.07	-	85.07	90.92	61.73	-	-	-	-	43	386
SSP [13]	67	79	-	-	41	-	9	-	-	194	977
Ours	**91.4**	**89.83**	**94.65**	**98.47**	**87.76**	10.63	**1.59**	25309	392	232	423

TABLE II

Results on the KITTI Tracking test set. For more details, visit HTTP://WWW.CVLIBS.NET/DATASETS/KITTI/EVAL_TRACKING.PHP

	MOTA	MOTP	MT	ML	IDS	FRAG
CIWT [21]	75.39	79.25	49.85	10.31	165	660
SCEA [23]	75.58	79.39	53.08	11.54	104	448
MDP [24]	76.59	82.10	52.15	13.38	130	387
LP-SSVM [3]	77.63	77.80	56.31	8.46	62	539
NOMT [6]	78.15	79.46	57.23	13.23	**31**	**207**
MCMOT-CPD [2]	78.90	82.13	52.31	11.69	228	536
Ours(RRC-IIITH)	**84.24**	**85.73**	**73.23**	**2.77**	468	944

TABLE III

Ablation Study. Comparision across various cues used for pairwise cost computation and choice of object detector. (App - appearance cost)

Cue(s)	MOTA	MOTP	Recall	Precision	MT	PT	ML	TP	FP	IDS	FRAG
SubCNN + App	66.18	82.60	86.45	89.29	71.63	24.46	3.90	23563	2825	708	1100
SubCNN + 3D-2D	68.24	82.60	86.45	89.29	71.63	24.46	3.90	23563	2825	429	829
SubCNN + 3D-3D	69.36	82.60	86.45	89.29	71.63	24.46	3.90	23563	2825	377	778
SubCNN + 3D-2D + App	70.96	82.60	86.45	89.29	71.63	24.46	3.90	23563	2825	472	870
SubCNN + 3D-3D + 3D-2D + App + Shape-Pose	**71.52**	**82.60**	**86.45**	**89.29**	**71.63**	**24.46**	**3.90**	**23563**	**2825**	**338**	**740**
RRC + App	80.53	89.83	94.62	98.47	87.76	10.63	1.59	25309	392	2863	3022
RRC + 3D-2D	86.91	89.83	94.65	98.47	87.76	10.63	1.59	25309	392	1328	1507
RRC + 3D-3D	87.56	89.83	94.65	98.47	87.76	10.63	1.59	25309	392	1170	1333
RRC + 3D-2D + App	89.65	89.83	94.65	98.47	87.76	10.63	1.59	25309	392	668	849
RRC + 3D-3D + 3D-2D + App + Shape-Pose	**91.4**	**89.83**	**94.65**	**98.47**	**87.76**	**10.63**	**1.59**	**25309**	**392**	**232**	**423**

TABLE IV

Results using Shape and Pose along with other Cues

	MOTA	MOTP	Recall	Precision	MT	PT	ML	TP	FP	IDS	FRAG
w/o Shape and Pose	55.3	86.01	98.9648	84.75	100	0	0	478	86	5	7
with Shape and Pose	**57.29**	86.01	98.9648	84.75	100	0	0	478	86	**1**	**5**

For example the first column of Fig 4 shows cars occluded on either sides of the road accurately tracked almost till the horizon. Whereas the second column shows efficient tracking of cars at varying depths and varying poses in an intersection while the third column shows precise tracking of occluding cars as well as a car that is being overtaken from the right by the ego car. In fact in the 4th frame a very small portion of the car is visible yet accurately tracked.

H. Summary of Results

The cornerstone of this effort is that single view monocular 3D cues obtained though formalisms developed on the basis of single view geometry can be effectively exploited to track vehicles in challenging scenes. This gets illustrated in the various tabulations of this section.

Table I depicts significant improvements over many of the current state of the art methods with a tracking accuracy in excess of 90%. We test our approach on the KITTI Tracking online server. Table II depicts significant improvements over the published approaches, with tracking accuracy over 84%.

Whereas the ablation studies in Table III does showcase the repeatability of 3D cues in improving the baseline appearance only tracking over detectors. While not as significant as in [15] baseline improvement over SubNN object detector[16] can be gleaned from Table III. The improvement in ID switches and fragmentations can also be seen over both detector baselines as a consequence of the 3D cues.

Table IV shows the relevance of pose and shape cues over a subsequence where association costs due to such cues improves baseline performance.

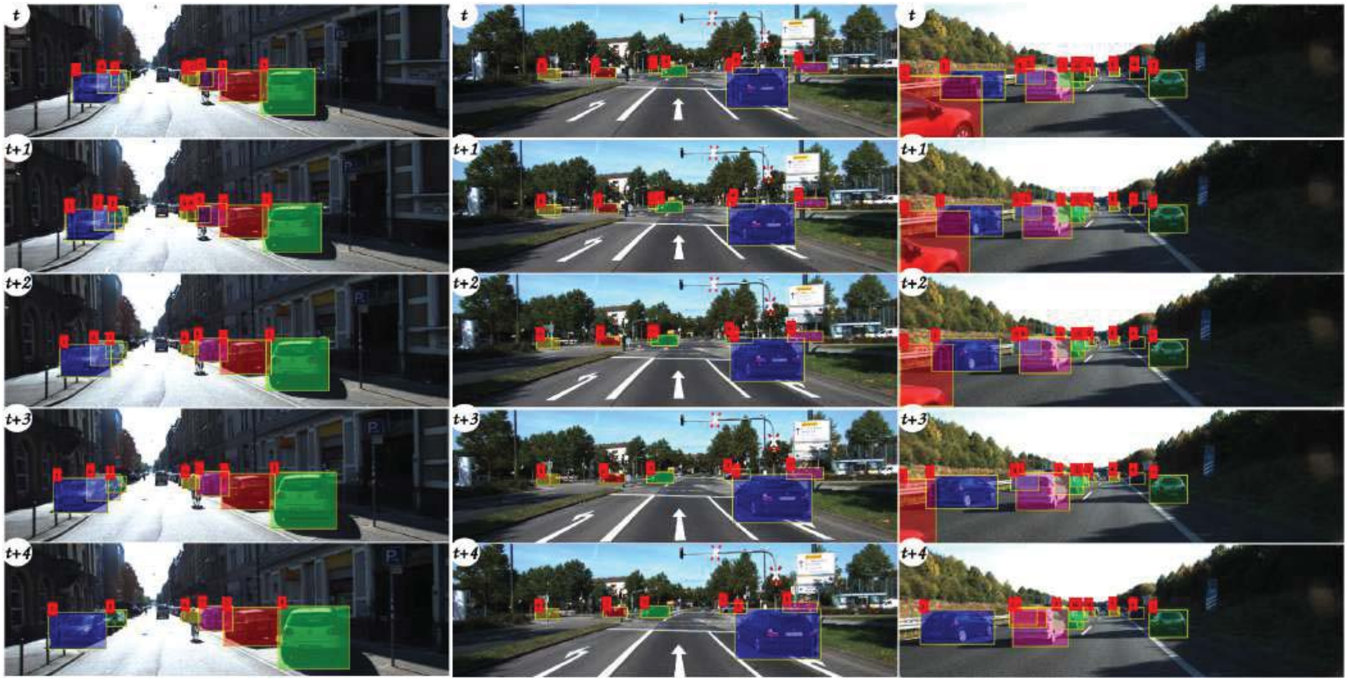

Fig. 4. Qualitative results on some challenging sequences.

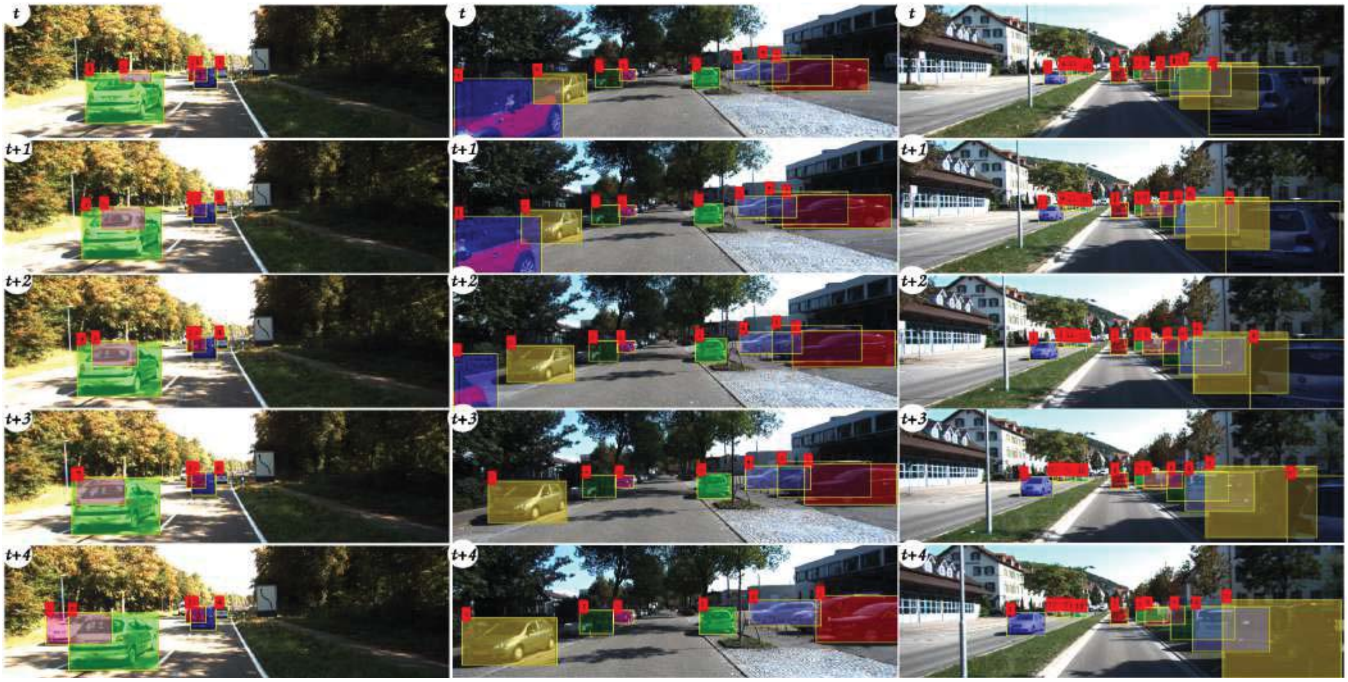

Fig. 5. Qualitative results on some challenging equences.

VI. Conclusions

Most state of the art tracking formalisms have not explored the role of 3D cues and when they have done those cues have been due to immediately available stereo depth. This paper showcased for the first time monocular 3D cues obtained from single view geometry along with pose and shape cues results in the best tracking performance on popular object tracking training datasets. These cues result in a set of simple, intuitive pairwise costs for multi-object tracking in a tracking-by-detection setting. Despite being more difficult to compute than ready made 3D depth data, monocular 3D cues have a role to play in diverse on road applications including object and vehicle tracking. Apart from the quantitative, qualitative results too signify its advantage in challenging scenes that involve considerable occlusions, minimal appearance of the object in the scene and objects that are far enough that they appear on the horizon. Although we demonstrated results using a simple Hungarian method based tracker, incorporation of sophisticated trackers would result in even higher performance boosts.

References

[1] S. Schulter, P. Vernaza, W. Choi, and M. Chandraker, "Deep network flow for multi-object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[2] B. Lee, E. Erdenee, S. Jin, M. Y. Nam, Y. G. Jung, and P. K. Rhee, "Multi-class multi-object tracking using changing point detection," in *European Conference on Computer Vision*. Springer, 2016.

[3] S. Wang and C. C. Fowlkes, "Learning optimal parameters for multi-target tracking with contextual interactions," *International Journal of Computer Vision*, 2017.

[4] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011.

[5] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM computing surveys (CSUR)*, 2006.

[6] W. Choi, "Near-online multi-target tracking with aggregated local flow descriptor," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[7] J. K. Murthy, G. S. Krishna, F. Chhaya, and K. M. Krishna, "Reconstructing vehicles from a single image: Shape priors for road scene understanding," in *Proceedings of the IEEE Conference on Robotics and Automation*, 2017.

[8] J. K. Murthy, S. Sharma, and M. Krishna, "Shape priors for real-time monocular object localization in dynamic environments," in *Proceedings of the IEEE Conference on Intelligent Robots and Systems (In Press)*, 2017.

[9] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[10] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008.

[11] A. Andriyenko, K. Schindler, and S. Roth, "Discrete-continuous optimization for multi-target tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012.

[12] A. Dehghan, S. Modiri Assari, and M. Shah, "Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[13] P. Lenz, A. Geiger, and R. Urtasun, "Followme: Efficient online min-cost flow tracking with bounded memory and computation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[14] W. Choi, C. Pantofaru, and S. Savarese, "A general framework for tracking multiple people from a moving camera," *IEEE transactions on pattern analysis and machine intelligence*, 2013.

[15] J. X. J. W. J. QiongYan and Y.-W. LiXu, "Accurate single stage detector using recurrent rolling convolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[16] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Subcategory-aware convolutional neural networks for object proposals and detection," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE, 2017.

[17] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn, "Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 120–127.

[18] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE transactions on pattern analysis and machine intelligence*, 2011.

[19] V. Chari, S. Lacoste-Julien, I. Laptev, and J. Sivic, "On pairwise costs for network flow multi-object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[20] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "Robust multiperson tracking from a mobile platform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1831–1846, 2009.

[21] A. Osep, W. Mehner, M. Mathias, and B. Leibe, "Combined image-and world-space tracking in traffic scenes," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017.

[22] S. Song and M. Chandraker, "Joint sfm and detection cues for monocular 3d localization in road scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[23] J. H. Yoon, C.-R. Lee, M.-H. Yang, and K.-J. Yoon, "Online multi-object tracking via structural constraint event aggregation," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[25] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, 2008.

[26] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics (NRL)*, 1955.