Temporal Semantic Motion Segmentation using Spatio Temporal Optimization

Nazrul Haque[1], Dinesh Reddy[2], and Madhava Krishna[1]

[1] International Institute of Information Technology, Hyderabad, India
[2] Robotic Institute, Carnegie Mellon University, Pittsburgh, USA

Abstract. Segmenting moving objects in a video sequence has been a challenging problem and critical to outdoor robotic navigation. While recent literature has laid focus on regularizing object labels over a sequence of frames, exploiting the spatio-temporal features for motion segmentation has been scarce. Particularly in real world dynamic scenes, existing approaches fail to exploit temporal consistency in segmenting moving objects with large camera motion.
In this paper, we present an approach for exploiting semantic information and temporal constraints in a joint framework for motion segmentation in a video. We propose a formulation for inferring per-frame joint semantic and motion labels using semantic potentials from dilated CNN framework and motion potentials from depth and geometric constraints. We integrate the potentials obtained into a 3D(space-time) fully connected CRF framework with overlapping/connected blocks. We solve for a feature space embedding in the spatio-temporal space by enforcing temporal constraints using optical flow and long term tracks as a least-squares problem. We evaluate our approach on outdoor driving benchmarks - KITTI and Cityscapes dataset.

1 INTRODUCTION

Understanding scene dynamics has always been a crucial component in outdoor robotic navigation. In outdoor scenes, scene understanding is facilitated by predicting spatially separated bounding boxes [19] [20] on objects or associating a *label* with each pixel [30] [1], in an image. For a holistic perception of the scene, the prediction unfolds in assigning a *semantic* or *motion* label. However, both the cues provide complementary information about a scene and are highly interrelated. Joint information about an object such as Moving Car or Stationary pedestrians significantly aids in path planning for an autonomous vehicle. Semantic property of an object can help infer the motion label of the pixel and vice versa.

In both static and dynamic environments, convolutional neural networks have gained enormous success in accuracy for predicting semantic labels in image space. On the other hand, motion segmentation poses many challenges, particularly in scenes where the camera is found to be in motion. Indeed, in the presence of multiple moving objects, generating and tracking various prospective motions becomes challenging. While epipolar geometry constraints work well with moving object detection, they tend to fail in degenerate cases where both the moving object and camera lie in the same subspace. This relative configuration between the camera and the object in motion is common in on-road scenes. Traditional

motion segmentation algorithms formulate the problem as clustering the trajectories into affine subspaces. This often leads to a sparse segmentation resulting in different clusters, each representing a motion model in the scene. Further, supervoxel projections on the trajectory clusters give rise to a dense segmentation. In most cases, the projections do not respect the object boundaries and hence, this is followed by a graph based probabilistic model such as Markov Random Field(MRF) to enforce appearance constraints.

Recent literature[**?**][18][17][29] have leveraged semantic property into a probabilistic framework, generating per-frame moving object proposals with dense predictions. Camera motion often leads to discontinuities in the flow magnitude. Optical flow magnitude for nearby stationary objects may be found to have a larger magnitude than far away stationary objects. This, in turn, effects the motion likelihood that is obtained using depth information in a similar fashion. Thus to eliminate failure cases semantic property comes into role. The intuition behind such a reasoning is that the likelihood of a moving wall or road is less as compared to a moving car or pedestrian.

In this paper, we focus on the problem of joint semantic and motion segmentation. Our method takes a sequence of stereo sequences as input and generates per frame motion probabilities using stereo pairs. For semantics, we use a dilated convolution neural network for predicting per-pixel semantic class. We also learn the correlation between the semantic label and motion likelihood. Further, We propose a novel joint probability formulation in a discrete label space consisting of joint labels. Under this, each image pixel is labeled with both semantic and motion class jointly. Motion property of an object is better perceived by the object tracks over the temporal space. To infer motion probabilities and enforce long term correspondences in image space, we use a dense fully connected CRF (Conditional Random Field) with time as an additional dimension in its feature space. The trajectory constraints are enforced by solving a linear least squares equation for optimizing position features in pairwise constraints in CRF.

The temporal constraints are enforced using dense point trajectories and optical flow. In addition, the spatial properties are preserved by including a second order regularization term in the least squares formulation. The spatial term exploits appearance similarity and edge maps for minimizing the distance between corresponding points in the sequence. For inference, we use an extension of the mean-field based algorithm. The inference is carried out on overlapping connected components in CRF.

In summary, following are the key contributions of our work.

- We present an *end-to-end probabilistic framework* that performs joint semantic and motion segmentation for a sequence of stereo frames.
- We provide a method for integrating semantic constraints with dense point object trajectories to obtain motion segmentation.
- We present results on several sequences on outdoor driving benchmarks.

We evaluate our approach on challenging KITTI on-road dataset. We are able to achieve 4.71% and 17.91% improvement in IOU accuracy on our annotated test dataset for Moving Car and Moving Pedestrian detection, respectively over M-CRF [18], while in comparison to STMOP[5], we show an improvement of 8.04% than M-CRF[18] in moving objects detection.

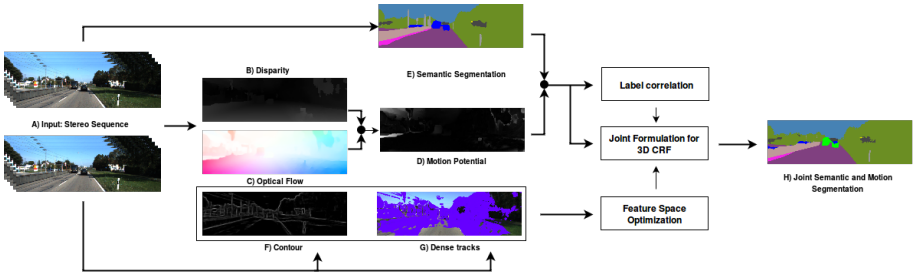

Fig. 1. Illustration of the proposed pipeline: Our framework takes a long sequence of stereo images as input (A). We compute motion potential(D) using depth(B) and Optical Flow(C). Semantic Segmentation(E) is carried out on the input images. Further, we calculate label compatibility between and object and motion class. We compute edge maps (H) and dense point trajectories (G) and solve a least-squares optimization for joint label CRF feature space enforcing temporal consistency and spatial constraints. Using the optimized feature space, we propose a joint CRF formulation in space-time volume. This leads to a temporally consistent joint semantic and motion segmentation (H).

2 RELATED WORK

Scene understanding has fair amount of literature in both semantic and motion aspects. Traditional semantic segmentation approaches have tackled the problem as a multi label classification problem. Classifiers are trained with descriptive features as input for dense pixel labeling[**?**], followed by a maximum a posteriori inference (MAP) in a conditional random field (CRF) [11] [18]. With the advent of convolutional neural networks(CNN), the dense pixel predictions have made significant progress on the accuracy[24]. Fully Convolution Networks [15] have made it possible for the architectures to handle inputs with arbitrary size. The outputs from the Convnets are upsampled by learning a Deconvolution layer [16], resulting in pixel wise predictions. The literature also includes architectures where Convnets followed by a CRF formulation [31] attain significant improvement in accuracy. Koltun [30] proposed an architecture for dense pixel predictions with dilated convolutions and presents state-of-the-art results in semantic segmentation. In recent work, Kundu[14] has shown results for temporally consistent semantic segmentation over a video sequence using long term tracks into a 3D CRF formulation. Existing architectures suffer loss in resolution due to pooling layers, while dilated convolutions sustain exponential expansion of the receptive field without loss in the coverage area. This also leads to a higher resolution output.

In outdoor scenes, motion segmentation has been extensively studied. Traditional algorithms based on epipolar geometry[27] are bound to fail in degenerate cases. The problem is tackled with good precision using geometric constraints [13], frame depth and camera ego-motion. In [13], degeneracy is handled by enforcing flow vector constraints using the camera trajectory obtained from visual SLAM (VSLAM). While the approaches demonstrate good accuracy on the on-road benchmark, they fail to exploit object trajectories for a consistent motion segmentation. In contrast, our approach uses dense point trajectories and semantic constraints in a joint framework for moving objects detection.

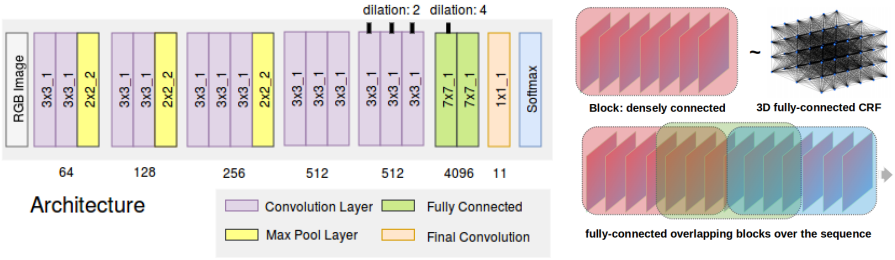

Fig. 2. (left)Dilated CNN architecture for semantic segmentation - $w \times h\_s$: Layer with kernels of width w, height h, and stride s. Numbers on the top and bottom of each layer depict the number of channels in the output and dilation factor, respectively. (right)3D CRF structure: The long sequence falls under overlapping blocks. For each block, a dense CRF is formulated and feature optimization is carried out for each set.

In other seminal works [26][10], the problem is formulated as affine clustering of the trajectories into corresponding subspaces. [10] [32] do not scale well on on-road datasets where both camera and the moving object lie in the same subspace. [26] used in-frame shear constraints for merging perspective affine models and has shown benchmark results on outdoor scenes. The output, however, is a sparse collection of points belonging to different subspaces based on the motion property exhibited. [5] used uncalibrated frame sequences for generating per frame moving object proposals. The proposals are refined by training dual pathway CNN with both the RGB image and optical flow as input. Further, labels are propagated using random walkers on motion affinities of long term tracks. While the approach works well with video segmentation benchmarks, they do not scale well on outdoor scenes. Our approach likewise exploits the long term tracks, while preserving semantic properties, and uses flow bound constraints for generating per-frame motion likelihood.

In recent literature, semantic properties have been exploited for motion segmentation. [3] uses contextual descriptors for object level motion detection, while semantics has been incorporated in a convolutional neural network architecture [29] for analyzing pedestrians behavior. With the advent of CNNs, efforts for learning joint labels in an end-to-end architecture has been studied in [8] using feature amplification, exploiting short term consistency. The idea has also been complemented by the work by Valada et al.[**?**] where joint learning is performed using two streams, each tasked to learn semantic and motion attributes. The two parallel streams are fused for joint learning and probability maps thus obtained are subsequently upsampled to obtain joint dense predictions. [18] [**?**] [8] fail to establish long term correspondences over a sequence for motion detection, whereas our approach uses long term tracks for establishing correspondences between the frames and enforce spatial constraint using appearance and edge features using the formulation.

3 Dynamic video joint segmentation

In this section, we present our joint labeling framework for outdoor sequences using a stereo camera. We describe our procedure for obtaining per-frame semantic potentials and motion initialization using stereo vision. We present our

joint formulation for dense CRF in space-time volume with overlapping connected blocks in the succeeding section. Further, we describe the least squares formulation in the joint label space for long term consistency, using dense point trajectories and spatial constraints.

3.1 Semantic Segmentation

For semantic class segmentation, we use a deep learning architecture specifically engineered for dense predictions. The architecture is adapted from fully convolutional VGG 16 net[15][23] and modifications applied from work by Yu and Koltun [30] using dilated convolutions. The last two pooling layers in the VGG architecture were detached and following convolutions are dilated with a factor of *2* for each pooling layer abducted. The dilated architecture benefits dense predictions by generating higher resolution output without losing global context. The network architecture used is shown in Fig. 2 The network takes full size color images as input and the output from the softmax layer is upsampled using a learned Deconvolution layer.

3.2 Motion Potential

We calculate per-frame motion likelihood using stereo pairs. The motion likelihood of a pixel is initialized as the difference between the predicted flow and optical flow vector. The camera extrinsics are calculated using libviso[7]. Given a stereo pair at consecutive time instants, the method estimates visual odometry by minimizing the sum of the reprojection errors on the both the images (left and right). The predicted flow vector of a pixel is stated as:

$$F^{'} = (KRK^{-1}X + \frac{KT}{z}) - X \tag{1}$$

where R and T are the rotation and translation of the camera respectively, K is the camera Intrinsic matrix, X is the pixel coordinate, z is the depth of the given pixel from the camera and $F^{'}$is the displacement of the pixel under the camera motion. The difference between the predicted and optical flow gives the motion potential for the pixel.

4 Joint labeling in space-time volume

In this section, we propose a space-time CRF formulation for joint semantic and motion labeling on a sequence of stereo frames. We also incorporate long term tracks in the joint feature space.

4.1 Spatio-Temporal CRF

Given a sequence of frames, we divide the video sequence into overlapping blocks and formulate a Fully Connected dense CRF on each block (Fig. 2). We extend the 3D CRF formulation of [14] to joint label space. In [14], position features of the CRF model are optimized using dense trajectories for generating temporally consistent semantic segmentation. Since motion segmentation is perceived better

with tracking objects over a sequence, we provide an extension of the framework for joint label space. We introduce the terms used in this paper. Each pixel in the video volume is located by the vector $\mathbf{p} = (n, t, i) \in \mathbb{R}^3$. Here, n is the block number, t denotes time dimension inside the block n, representative of the frame number relative to the block and i is the index of the pixel in the image. $\mathbf{P}$ represents a group of pixels in the volume. The location of a pixel $\mathbf{p}$ is given by $\bar{x}_p$ in the image space. Also, the RGB color vector of a pixel $\mathbf{p}$ is represented by $\mathbf{C}_p$.

For a given block in the video volume, we define the problem of joint semantic and motion segmentation as finding a minimal cost labeling on a the joint label space $\mathcal{L} = \{l_1, l_2, ..., l_k\}$ for a set of random variables $\mathcal{X}_p = \{x_1, x_2, ..., x_\mathcal{N}\}$. We denote s as the total number of semantic class labels. Each random variable can take a single label from joint label space $\mathcal{L}$, where $k = s * 2$, as each object class label can be associated with two motion classes (moving or stationary). For instance, l_i could be a moving car, stationary road, etc. The energy cost term for a label assignment x is defined as:

$$E(x|\mathbf{P}) = \sum_i \psi_i\left(x_i\right) + \sum_{(i,j)\mathcal{N}} \psi_{i,j}\left(x_i, x_j\right) \tag{2}$$

where $\mathcal{N}$ is the neighborhood constitution of the random field defined on the pairs of variables. In the random field, a clique covers each block and each pixel is covered by two overlapping blocks. Subsequently, each variable falls under two fully connected subgraphs.

The unary term $\psi_i\left(x_i\right)$, represents the cost of assigning label x_i to pixel $\mathbf{i}$. For joint labeling, we propose the unary cost formulation as:

$$\psi_i\left(x_i\right) = -\log\left(\bar{\phi}_i\left(x_i\right)\right) \tag{3}$$

$$\phi_i\left(x_i\right) = \underbrace{\phi_{i,s}\left(x_i\right)}_{Object} . \underbrace{\phi_{i,m}\left(x_i\right)}_{Motion} . \underbrace{\phi_{i,s,m}\left(x_i\right)}_{Correlation} \tag{4}$$

Here, $\bar{\phi}_i\left(x_i\right)$ is obtained after normalizing the joint probability distribution $\phi_i\left(x_i\right)$ in the range $0 - 1$. $\phi_{i,s}\left(x_i\right)$ is the probability of the pixel belonging to the object class s corresponding to the joint label x_i and inferred from the probabilities obtained from the softmax layer in our trained dilated ConvNet described in section 3.1. We express the motion term as:

$$\phi_{i,m}\left(x_i\right) = \begin{cases} \lambda & if\ \bar{l}\left(m\right) = 0 \\ ||F^{'}\left(\bar{x_i}\right) - F(\bar{x_i})|| & if\ \bar{l}\left(m\right) = 1 \end{cases} \tag{5}$$

where, m is the motion attribute in the joint label. The function $\bar{l}(m)$ returns 1 if the motion label space is moving and 0 otherwise. For instance, $\bar{l}(m)$ will return 1 in case of 'Moving Car'. F' is the predicted flow(Sec 3.2) and F, the optical flow vector of the pixel i. $||F^{'}\left(\bar{x_p}\right) - F(\bar{x_p})||$ represents the normalized motion potential corresponding to the pixel. λ is a learned term calculated using RANSAC algorithm over a small set of annotated images.

$\phi_{i,s,m}\left(x_i\right)$ represents the object class label motion compatibility and is given as:

$$\phi_{i,s,m}\left(x_i\right) = c(s, m) \tag{6}$$

where, $c(s,m) \in [0,1]$. Here $c(s,m)$ represents the correlation between the semantic class s and motion attribute m. In other words, this can be seen as the motion compatibility for semantic class s. We follow the work of [18] for calculating label correlation between the two classes using MAHR algorithm [**?**].

The pairwise term $\psi_{i,j}\left(x_i, x_j\right)$ stimulates similar pixels to have same labeling. With Gaussian kernels, the pairwise term [11] given as:

$$\psi_{i,j}\left(x_i, x_j\right) = \mu\left(x_i, x_j\right) \sum_{z=1}^{Z} \omega^z \kappa^z\left(\mathbf{f}_i, \mathbf{f}_j\right) \tag{7}$$

Here, $\mathbf{f}_i$ and $\mathbf{f}_j$ are features analogous to **i** and **j** pixel respectively. $\mu\left(x_i, x_j\right)$ is a label compatibility term and ω^z are the weights. The kernels κ^z is given as:

$$\kappa^z\left(\mathbf{f}_i, \mathbf{f}_j\right) = \exp\left(-\frac{||\mathbf{f}_i - \mathbf{f}_j||}{\sigma_z^2}\right) \tag{8}$$

where, σ_z is the model parameter learned using grid search on a subset of annotated training set. The feature space $\mathbf{f}_i$ is a six dimensional vector $\in \mathbb{R}^6$ that consists of position, color and time corresponding to the pixel **i**, i.e, $\mathbf{f}_i = (\bar{x}_i, \mathbf{C}_i, t_i)$.

4.2 Feature Space Optimization

The six dimensional feature space do not scale well in dynamic outdoor scenes as the pixels tend to displace under the camera and object motion. Thus, time as an additional term in the feature space does not model the pixel correspondences in the space-time volume. We use the formulation proposed in [14] for optimizing position features ($\bar{x}_i$) to reduce the Euclidean distance between the corresponding points in the volume while enforcing spatial constraints for preserving object shapes. In [14], the position features were optimized for temporally consistent semantic segmentation using long term tracks and edge maps. However, we use the underlying formulation for enforcing temporal consistency in joint semantic and motion label space. The dense point trajectories and spatial constraints enforce label similarity with motion potential in unary space. Thus, for the feature space $(\bar{x}_i, \mathbf{C}_i, t_i)$, the position features($\bar{\mathbf{x}}_i$) are optimized using a least squares formulation. The optimized feature space is represented as $(\mathbf{x}_i, \mathbf{C}_i, t_i)$ which is obtained after minimizing the proposed energy term.

We use the linear least squares formulation in [14] and is given as:

$$E^{\mathcal{SM}}(x) = E_d^{\mathcal{SM}}(x) + E_s^{\mathcal{SM}}(x) + E_t^{\mathcal{SM}}(x) \tag{9}$$

where x are the position features in the block n, consisting of R frames, with N pixels in each frame. We now explain each of the terms in the energy equation 9. A single pixel is denoted by (n,t,i) as described in Sec. 4.1, with n as the block number.

Data $E_d^{\mathcal{SM}}(x)$: Let $r = \lfloor R/2 \rfloor$ be the reference frame. The energy term prevents the points in the reference from drifting far from their original position in the volume. If P^r is the set of pixels in the reference frame and $\bar{x}_p$ be the original feature space, the term is given as:

$$E_d^{\mathcal{SM}}(x) = \sum_{p \in P^a} \left(\mathbf{x}_p - \bar{x}_p\right)^2 \tag{10}$$

Spatial $E_s^{\mathcal{SM}}(x)$: This spatial energy term ensures that object shapes are preserved with color and boundary constraints. This is formulated over the 4-connected pixel grid and given as:

$$E_s^{\mathcal{SM}}(x) = \sum_{t=1}^{R} \sum_{i=1}^{N} \left(\mathbf{x}_{(n,t,i)} - \sum_{j \in \mathcal{N}_\rangle} \omega_{ij} \mathbf{x}_{(n,t,i)} \right)^2 \tag{11}$$

where, $\mathcal{N}_i$ are the neighbors of the point (n,t,i) The weights ω_{ij} reduces the regularization effect at boundaries for preserving object shapes and given as:

$$\omega_{ij} = \exp\left(- \frac{||C_{(n,t,i)} - C_{(n,t,j)}||}{\sigma_1} \right) \exp\left(- \frac{q_p}{\sigma_2} \right) \tag{12}$$

Here, q_p is the contour strength of the pixel $\mathbf{p}$ - (n,t,i). Hence, the second term in equation 12 is related to the contour strength at the pixel, while the initial term is based on the color difference between the pixels (n,t,i) and (n,t,j). $q_p \in [0,1]$, where higher value indicates presence of an edge at the pixel. For calculating contour strength q_p, we use Structured Forests Edge Detection [4] implementation.

Temporal $E_t^{\mathcal{SM}}(x)$: The energy term uses correspondences obtained from the dense point trajectories and optical flow for bringing corresponding points closer in the feature space and is given as:

$$E_t^{\mathcal{SM}}(x) = \sum_{(p,q) \in \mathcal{Y}} \left(\mathbf{x}_p - \mathbf{x}_q \right)^2 \tag{13}$$

where, $\mathcal{Y}$ is the super set of corresponding points in the frames of the block n. Here, points p and q belong to different frames. The energy term ensures that the tracked points over the frames are assigned the same label over the joint label space. This also enforces label compatibility for pixels exhibiting similar motion behavior over the frames. We use the implementation by Sundaram et al. [25] for calculating long term tracks.

4.3 Inference

Inference has been a challenging problem for dense CRFs and becomes even more challenging with a set of overlapping Fully Connected blocks. We follow the work of Kundu et al. [14], an extension of mean-field inference algorithm by Koltun [11]. Since a pixel is covered by two overlapping blocks, let $\mathcal{N}_i^1$ and $\mathcal{N}_j^2$ represent the two set of neighbors of the pixel $\mathbf{i}$. We define an alternative distribution over the random variables of the CRF, $Q_i(z_i)$, and define Q as $Q(z) = \prod_i Q_i(z_i)$. Here $Q_i(z_i)$ is a multi class distribution over the joint semantic and motion label space. The mean-field approach minimizes the distance between the Q and the true distribution P. The inference for joint label space is given as:

$$\begin{aligned} Q_i(x_i = l) = & \frac{1}{Z_i} \exp\Bigg(- \psi_i(x_i) - \sum_{l' \in \mathcal{L}} \sum_{j \in \mathcal{N}_i^1} Q_j(x_j = l') \psi_{ij}(l, l') \\ & - \sum_{l' \in \mathcal{L}} \sum_{j \in \mathcal{N}_i^2} Q_j(x_j = l') \psi_{ij}(l, l') \Bigg) \end{aligned} \tag{14}$$

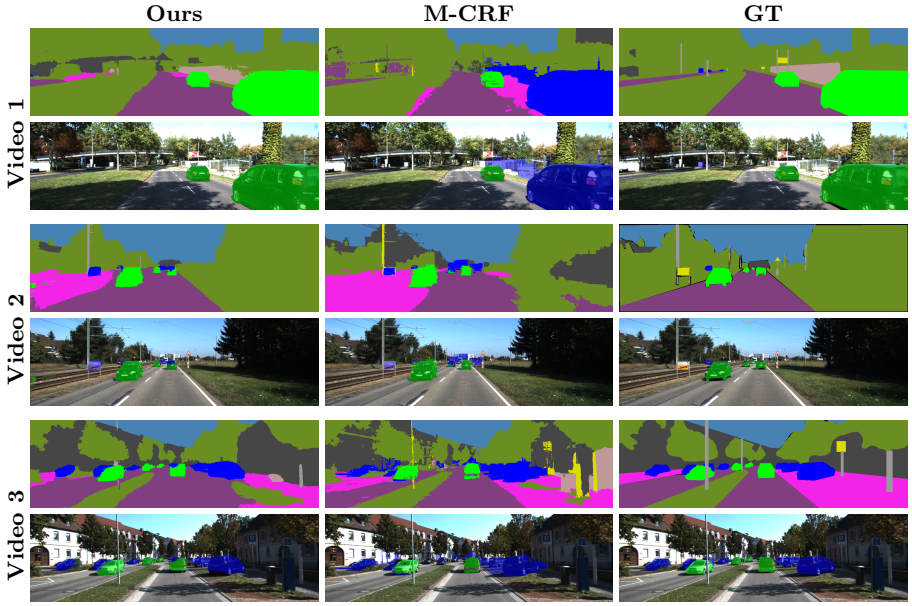

Fig. 3. Qualitative evaluation of our approach against M-CRF[18] and Ground Truth Labeling on our KITTI test dataset. The color convention of our joint labeling can be referred from Table 1. Left To Right: (1) Joint segmetation results from our approach on KITTI sequences. We also show overlay image for better visualization, where moving and stationary cars are overlayed by *green* and *blue* colour respectively. (2) Output from Multi Layer CRF [18] (3) Ground Truth annotations.

where, $\psi_i(x_i) = \psi_{i,s}(x_i) . \psi_{i,m}(x_i) . \psi_{i,s,m}(x_i)$ and Z_i is the normalization factor. As proposed in [11] we can efficiently solve the pairwise summations, given as Potts model, using Gaussian convolutions. In cases where blocks do not fit into the memory, inference is carried out in chunks and predictions are scaled across the divisions using the heuristic from the work by [14].

5 Evaluations and Results

For evaluation of our approach, we use two renowned on-road datasets.

KITTI: We use the KITTI-Tracking benchmark dataset [6] which consists of diverse on-road sequences taken by a stereo camera, mounted on a driving car. For quantitative evaluation, we use the largest publicly available annotated dataset (200 images) by [8]. The images were annotated with 11 semantic classes, i.e, *Building, Vegetation, Sky, Car, Sign, Road, Pedestrian, Fence, Pole, Sidewalk* and *Cyclist*. Further, each image was also annotated with moving and non-moving labels. We use the results provided by [18] on 200 images for both quantitative and qualitative evaluation, which is a subset of frames provided by [8] and we manually annotate the remaining images. Thus, we form our KITTI-Test dataset consisting of 200 images from five different sequences of KITTI Tracking

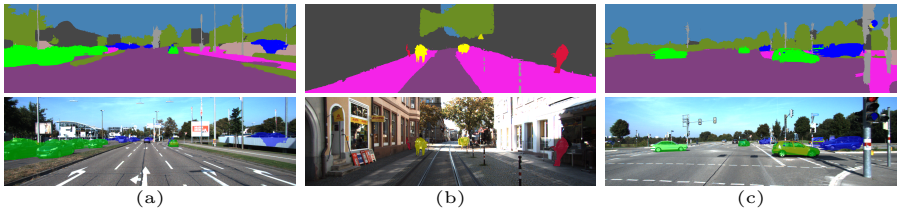

Fig. 4. Qualitative Evaluation on KITTI dataset. The results show proficiency of our approach across diverse scenes with challenging conditions.

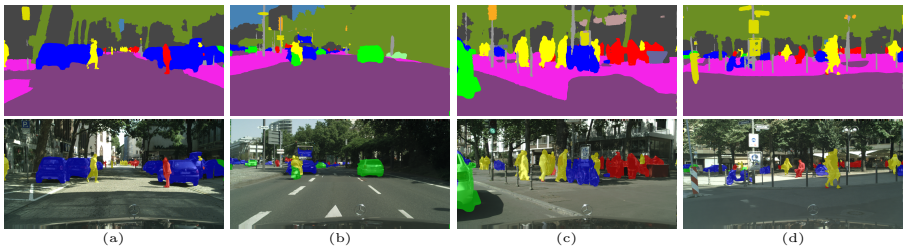

Fig. 5. Qualitative Evaluation on the CityScapes dataset. The joint segmentation obtained depict robustness of our approach with various dynamic objects.

dataset. Existing semantic segmentation annotations for KITTI dataset is insufficient to train a dilated CNN from scratch. Hence for fine-tuning our network, we manually annotate 56 images from KITTI sequences and ensure no overlap of sequences/images with our KITTI-Test dataset. The annotated images together with 146 KITTI Odometry images labeled by Ros et al. [21] forms our KITTI-Training dataset.

Cityscapes: The Cityscapes dataset [**?**] is relatively new and consists of challenging urban sequences from over 50 cities with varying dynamic objects and weather conditions. Pixel-wise ground truth semantic annotations are publicly provided for 2975 training and 500 validation images, for a single image in each video snippet. However, we use the semantic annotations for training our dilated CNN for the task of semantic segmentation. We show qualitative results obtained using our approach for joint segmentation on the video sequences provided in the validation set.

Training: We train our network for 20,000 iterations on Cityscapes training annotated set, with learning rate as 10^{-5}. Thereafter, the network was fine-tuned for 10,000 iterations on our KITTI-Training dataset, with learning rate and momentum as 10^{-4} and 0.9 respectively. We use this trained network for obtaining semantic segmentation for KITTI sequences. For obtaining semantic prediction for Cityscapes video sequences, we use the pre-trained dilated CNN network provided by [30]. The model was trained on Cityscapes training annotated benchmark with 19 semantic classes.

For computing disparity, we use Semi Global block matching algorithm [9]. We use state-of-the-art DeepFlow [28] for Optical Flow computation. The motion compatibility term is learned using our manually annotated training dataset. In the following section, we show an extensive qualitative and quantitative evaluation of our approach.

Qualitative Evaluation : We qualitatively evaluate our approach with respect to Ground Truth and M-CRF [18]. Figure 3 shows qualitative comparisons on video sequences for our KITTI-test dataset. The label color spectrum is consistent with the label descriptions given in table 3 and table 1. Particularly, *blue* and *green* colors denote static and moving car respectively. In Fig. 3, video sequence 1, we able to segment moving car approaching from behind while M-CRF [18] gives the stationary label to the object. M-CRF tends to rely on optical flow and in cases where a new object appears, optical flow is found to be inconsistent. With the incorporated temporal consistency in our approach, we are able to identify motion attributes for incoming objects in the scene. This is also evident in the video sequences 2 and 3 with multiple moving cars. In the sequences, M-CRF misses out on moving objects which are relatively far from the camera due to the inconsistent disparity in those regions. We are able to identify motion relatively farther from the camera which reiterates the role of a temporally consistent framework.

Temporal optimization in a joint label space proves advantageous in many ways. M-CRF relies on consistency constraints enforced through pairwise potentials in CRF and is found to have incorrectly labeled patches on the moving cars. With temporal optimization in joint label space, our approach enforces spatial coherence. There is also clear demarcation between the boundaries of the moving object and its surroundings in contrast to M-CRF [18]. This is attributed to the spatial constraints enforced in the least-squares formulation in combination with semantic segmentation. While M-CRF emphasizes a strong motion prior with a separate layer for semantics, our approach enforces consistent segmentation with long term tracks and integration with semantics in unary space. Also, the strong semantic prior obtained from the dilated CNN produces better object labeling.

Results across diverse scenes: To showcase the proficiency of our approach in diverse scenes and different moving objects, we show results on various KITTI-Tracking and Cityscapes sequences. We show both dense joint segmentation results and overlay-ed images with vehicle and pedestrian classes. The joint results are shown in Fig. 4, Fig. 5 and on complete sequences in the supplementary video `https://youtu.be/6kq8_FgwYFA`. Figure 4 (a) and (c) show highway scenes which consist of multiple cars moving with high speed. Our approach is correctly able to segment fleet of cars moving with high velocities (a). In Figure 4 (c) moving cars at the turns are clearly distinguished from nearby static cars in the scene, owing to the spatial constraints imposed while performing joint optimization. Due to the limited amount of annotated data for fine-tuning dilated CNN for KITTI dataset, the pedestrian classes are not identified with good precision across all sequences of KITTI. However, in sequences where semantic prediction for pedestrians is accurate, as shown in Figure 4(c), we are able to capture the motion behavior accurately using our approach.

Results with different moving agents: To show performance on different dynamic objects such as pedestrians, bikes, etc, we show qualitative results on Cityscapes dataset. The Cityscapes dataset proves beneficial in this regard due to its large semantic segmentation training dataset. The color codes in our output are consistent with the label spectrum given in table 3 and Cityscapes official color codes[**?**] for the remaining static classes. For better visualization, all vehicles are clubbed under a single class where moving and static vehicles are shown with *green* and *blue* color respectively. Similar policy is followed for 'human' classes where moving and static person is shown with *dark yellow* and *red* color

respectively. Also the void or unlabeled regions belonging to the camera mounted car are taken as road. Figure 5 (c) and (d) showcase our results on urban crowded scenes. We are able to segment moving pedestrians from the stationary ones with high precision and accurate boundaries which showcase the utility of our joint temporal optimization along with the spatial boundary constraints. The accurate differentiation can also be seen with the diverse vehicle classes present in the figure. The results are complemented in Figure 5 (a) which portrays a more common on-road urban street scene with human obstacles. In Figure 5 (a) our approach is correctly able to differentiate between on-road moving and stationary pedestrian as well as identify the motion attributes of pedestrians at a relatively larger distance from the camera. Figure (d) shows a person rested on a moving bike and it can be seen that our framework is able to categorize both of the agents as moving with precise boundaries. The results show the effectiveness of our approach in handling various classes across diverse scenes.

Table 1. (Left) Joint Motion Segmentation evaluation. We compare our method with Multi layer CRF [18]

Model	**Moving**		**Stationary**	
	Car	Pedestrian	Car	Pedestrian
M-CRF	68.87	19.27	28.72	16.08
Ours	**73.58**	**37.18**	**45.57**	**28.30**

Model	**Stationary**	**Moving**
STMOP	86.58	51.53
M-CRF	88.89	81.53
Ours-M	**96.94**	**89.17**

Table 2. (Right) Motion Segmentation evaluation on two annotated highway sequences. We compare our method with STMOP [5] and Multi layer CRF [18]

Quantitative Evaluation Quantitative evaluations are carried out with respect to M-CRF [18] and STMOP [5] which have been evaluated on video motion segmentation. For semantic labeling, we show our evaluation for semantic label space w.r.t to M-CRF [18], semantic CNN [30] and T-CRF [14].

Semantics: For quantitative evaluation of our semantic segmentation obtained after joint formulation, we compare our results with existing M-CRF , dilated CNN [30] and temporal semantic CRF [14]. Evaluation is staged by cross-verifying each pixel with the corresponding Ground Truth label. The evaluation metric used is intersection over union, defined as $TP/(TP+FP+FN)$, where TP represents True Positive, FP False Positive and FN False Negative over each pixel in the image. We use the KITTI-test dataset for quantitative evaluation of our approach. Table 1 shows quantitative evaluation in joint label space. Although we cannot see huge improvements over the semantic temporal CRF to our method, this can be attributed to the fact that label transfer using this method on dynamic and static object perform the same way. Our method shows an improved segmentation of the moving objects.

Motion: We show quantitative evaluation of our motion segmentation with respect to existing approaches in both stereo(M-CRF) and monocular setup STMOP. For a fair comparison with STMOP, we use a subset of our annotated test sequences consisting of relatively fewer moving cars. Also, we use the best supervoxel projection in the proposals generated by STMOP. Table 2 shows

quantitative evaluation of our approach. The main contribution of the proposed method can be seen in the Table 1. We observe a clear improvement in motion segmentation compared to the older methods which combined semantics and motion as a joint problem and geometric based methods. Most of the previous methods have attempted motion segmentation using geometric cues but could not get high accuracy because of the constraints in geometry. We show that combining learning based methods to geometric constraints can boost the accuracy of motion segmentation.

Method	Building	Vegetation	Sky	Car	Sign	Road	Pedestrian	Fence	Pole	Sidewalk
Multi Layer CRF [17]	43.56	65.41	70.01	71.17	2.06	59.29	39.2	50.40	9.71	31.96
Dilated [30]	59.38	83.16	**91.41**	82.17	**12.59**	83.72	**65.74**	52.69	**42.01**	41.29
Kundu et al. [14]	60.95	**83.41**	91.11	**82.63**	6.96	84.63	64.81	**53.03**	20.54	42.22
Ours	**61.20**	83.18	91.38	80.87	3.11	**84.69**	64.32	53.01	17.81	**43.87**

Table 3. Quantitative evaluation semantic segmentation with respect to M-CRF[18], dilated [30] and [14] on our annotated KITTI dataset.

6 Conclusion

We presented an end-to-end framework for joint semantic motion segmentation on a video. The proposed method integrates semantic constraints with dense point correspondences. We show results on multiple sequences of KITTI and release our annotations for comparisons.

We look at the problem of dynamic scene understanding and approach the problem using graphical models as end-to-end neural networks for these tasks need stronger cues. Looking at the end-to-end model for complete dynamic scene understanding is still an open problem. We believe that the dataset released and the current approach can form a basis for such models.

References

1. Badrinarayanan, V., Handa, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv preprint arXiv:1505.07293 (2015)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014)
3. Chen, T., Lu, S.: Object-level motion detection from moving cameras. IEEE Transactions on Circuits and Systems for Video Technology (2016)
4. Dollár, P., Zitnick, C.L.: Fast edge detection using structured forests. In PAMI (2015)
5. Fragkiadaki, K., Arbeláez, P., Felsen, P., Malik, J.: Learning to segment moving objects in videos. In: CVPR. IEEE (2015)
6. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR (2012)
7. Geiger, A., Ziegler, J., Stiller, C.: Stereoscan: Dense 3d reconstruction in real-time. In: Intelligent Vehicles Symposium (IV) (2011)

8. Haque, N., Reddy, D., Krishna, M.: Joint semantic and motion segmentation for dynamic scenes using deep convolutional networks. In VISAPP (2017)
9. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Transactions on PAMI (2008)
10. Jain, S., Madhav Govindu, V.: Efficient higher-order clustering on the grassmann manifold. In: ICCV. pp. 3511–3518 (2013)
11. Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. NIPS (2011)
12. Komodakis, N., Paragios, N., Tziritas, G.: Mrf energy minimization and beyond via dual decomposition. IEEE transactions on PAMI 33(3), 531–552 (2011)
13. Kundu, A., Krishna, K., Sivaswamy, J.: Moving object detection by multi-view geometric techniques from a single camera mounted robot. In: IROS. pp. 4306 –4312 (2009)
14. Kundu, A., Vineet, V., Koltun, V.: Feature space optimization for semantic video segmentation. In: CVPR (2016)
15. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: ICCV. pp. 3431–3440 (2015)
16. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: ICCV. pp. 1520–1528 (2015)
17. Reddy, N.D., Singhal, P., Chari, V., Krishna, K.M.: Dynamic body vslam with semantic constraints. In: IROS 2015
18. Reddy, N.D., Singhal, P., Krishna, K.M.: Semantic motion segmentation using dense crf formulation. In: ICVGIP (2014)
19. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR (2016)
20. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS (2015)
21. Ros, G., Ramos, S., Granados, M., Bakhtiary, A., Vazquez, D., Lopez, A.: Vision-based offline-online perception paradigm for autonomous driving. In: WACV (2015)
22. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. International Journal of Computer Vision 81(1), 2–23 (2009)
23. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS. pp. 568–576 (2014)
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
25. Sundaram, N., Brox, T., Keutzer, K.: Dense point trajectories by gpu-accelerated large displacement optical flow. In: ECCV. pp. 438–451. Springer (2010)
26. Tourani, S., Krishna, K.M.: Using in-frame shear constraints for monocular motion segmentation of rigid bodies. Journal of Intelligent & Robotic Systems 82(2), 237–255 (2016)
27. Vidal, R., Sastry, S.: Optimal segmentation of dynamic scenes from two perspective views. In: CVPR. vol. 2 (2003)
28. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: Deepflow: Large displacement optical flow with deep matching. In: ICCV (2013)
29. Yi, S., Li, H., Wang, X.: Pedestrian behavior understanding and prediction with deep neural networks. In: ECCV. pp. 263–279. Springer (2016)
30. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
31. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: ICCV. pp. 1529–1537 (2015)
32. Zografos, V., Nordberg, K.: Fast and accurate motion segmentation using linear combination of views. In: BMVC (2011)