MoA-Net: Self-supervised Motion Segmentation

Pia Bideau(✉), Rakesh R. Menon(✉), and Erik Learned-Miller(✉)

College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, USA
{pbideau,rrmenon,elm}@cs.umass.edu

Abstract. Most recent approaches to motion segmentation use optical flow to segment an image into the static environment and independently moving objects. Neural network based approaches usually require large amounts of labeled training data to achieve state-of-the-art performance. In this work we propose a new approach to train a motion segmentation network in a self-supervised manner. Inspired by visual ecology, the human visual system, and by prior approaches to motion modeling, we break down the problem of motion segmentation into two smaller subproblems: (1) modifying the flow field to remove the observer's rotation and (2) segmenting the rotation-compensated flow into static environment and independently moving objects. Compensating for rotation leads to essential simplifications that allow us to describe an independently moving object with just a few criteria which can be learned by our new motion segmentation network - the *Mo*tion *A*ngle *Net*work (MoA-Net). We compare our network with two other motion segmentation networks and show state-of-the-art performance on Sintel.

Keywords: Optical flow · Motion segmentation
Video segmentation · Camera motion · Visual ecology

1 Introduction

The human visual system has an incredible ability to detect motion, regardless of its complexity. While we are moving through the world our eye captures an enormous number of images over time. Images are projected onto our retina and the perceived motion (image change over time) is processed by the brain. In computer vision, optical flow is used to describe the motion between two consecutive images. Low level optical flow methods are based on two images alone [4,6,15,22,26,27]; other methods attempt to incorporate object knowledge and the knowledge about object motions [7,25,36]. In this work we propose a new approach to learning motion segmentation given an optical flow field as input.

The task of motion segmentation attempts to analyze the perceived motion and to segment a video sequence into the static environment (if any) and independently moving objects. Interpreting the motion field accurately and then

© Springer Nature Switzerland AG 2019

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11134, pp. 715–730, 2019.
https://doi.org/10.1007/978-3-030-11024-6_55

drawing the right conclusions about what is moving in the world and what is static is a complex process. To get a sense for the complexity of the task of motion processing in the brain, we consider three different situations that produce very different optical flow fields.

The first situation pictures a stationary scene (with no camera motion) in which one object is moving. This might be a person walking in the world, which is pictured as a person moving across the observer's retina. This case is simple to interpret. The perceived motion on the retina exactly corresponds to the motion in the world.

The second situation pictures a stationary scene in which the observer is turning his head (rotating), walking through the world (translating), or rotating and translating at the same time. If the observer is only rotating, the entire image moves across the retina according to the observer's motion. If the observer is translating, the pictured motion on the retina is far more complex. The perceived motion depends on the scene geometry. Objects that are close lead to a "faster" motion than farther objects. Objects at the horizon create no change on the retina. Observing these different types of motion on the retina can be interpreted in several ways: (1) the entire world is moving while the observer stands still, (2) different "speeds" of motion might lead to the conclusion that some objects are moving faster than their environment or alternatively, that objects might be located at different depths, (3) the observer moves while the world is standing still. In many cases the last option may appear to be the most reasonable interpretation of the observed motion on the retina.

The third and last situation to consider is scene with both a moving observer and moving objects. Both the observer's motion and the object motions result in motion of the scene pictured on the retina as described in the previous two cases. However, if we track a moving object with our eyes, the object will not create any motion on the retina. The object will appear to be stationary while the world appears to be moving.

These three situations show that just because something is moving across the retina does not mean that it is actually moving in the world. *How do humans know what is moving in the world and what is not?* This question is the subject of current research in many different areas such as neuroscience, psychology, and computer science.

Eye movements play a key role in simplifying optical flow on the eye's retina and making it easier to interpret for our brain [33]. Optical flow produced by eye motions (rotations) contains no information about the scene's geometry and thus can be used for motion compensation without adding or reducing critical information. The two major reasons for eye movements are (1) to stabilize vision and (2) to change direction of gaze.

In this work we aim to develop an approach that accurately interprets the perceived optical flow on the retina. Inspired by visual ecology, we start with vision stabilization before processing the optical flow to segment independently moving objects. Of course we are not able to receive an image directly from our

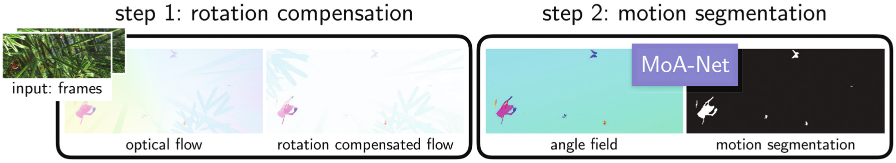
step 1: rotation compensation
step 2: motion segmentation
input: frames
optical flow
rotation compensated flow
MoA-Net
angle field
motion segmentation

Fig. 1. Self-supervised Motion Segmentation. Given an optical flow our goal is to segment a frame into independently moving objects and static environment. Due to the complexity of optical flow fields, previous neural network models have had difficulty segmenting motion directly from optical flow. As in previous work [1], we use a two step approach, which first involves adjusting the optical flow for camera rotation (left) and then segments the angle of the compensated flow into static environment and moving objects (right). By training a network to segment from the angle field rather than raw optical flow, we significantly improve performance.

eye. Instead we use video sequences taken by a camera and methods to estimate the optical flow between two consecutive frames.

Unlike most CNN-based approaches, we are not relying on labeled training data, which is limited. Instead we carefully analyse the underlying geometry of optical flow and break down the problem of motion segmentation into two subproblems: compensating the optical flow for rotation (similar to vision stabilization of our eye movements) and segmenting the remaining optical flow into static background and moving objects. The step of compensating the flow for camera rotation is a challenging step especially since flow field is a noisy estimate of the motion field [1,2]. Estimating the camera rotation given the optical flow as input is not further explored in this work; approaches to estimate the camera rotation are presented in [1,2].

As stated already in several previous literature [1,2,8,17] the remaining optical flow (after compensating for rotation) has a simple geometrical pattern. We use this simple geometrical pattern of rotation compensated flow to synthesize training data in large amounts following rules of perspective projection. In this way we do not rely on any training dataset for motion segmentation, which are limited in size, the variety of shown scene structures, or quality.
Our contributions are as follows:

- Inspired by visual ecology, we present a two step approach for motion segmentation, which first involves compensation of the optical flow for camera rotation and then segments the compensated flow into static environment and independently moving objects. While this two step approach is a well established approach for motion segmentation [1,2,8], we present in this paper its great potential for *learning based* video segmentation methods. We aim to leverage the strength of classical geometrical approaches (based on perspective projection) and learning based approaches for motion segmentation.
- For evaluation purposes, motion segmentation ground truth for Sintel is generated and will be made publicly available.

- A new self-supervised training approach is presented that does not rely on limited training data. Instead the problem of motion segmentation is broken down into two smaller subproblems. Guided by perspective projection, we provide a highly simplified (abstract) definition of a moving object, which allows us to generate an unlimited amount of training data in a synthetic way.
- We show state-of-the-art performance on ground truth optical flow of Sintel.

Our paper is organized as follows. In Sect. 3 we review the flow field and how scene geometry, object motion and observer motion contribute to the formation process of optical flow. This geometrical background information leads us to a new approach of training a neural network, which is described in Sect. 4. Rather than relying on labeled training data, synthetic training data is automatically generated considering the geometry of optical flow. In Sect. 5 we evaluate our motion segmentation network and compare its performance to two other recently published networks for motion segmentation.

2 Related Work

Motion segmentation is studied for many years in the area of computer science [24,31,34] as well as neuroscience and psychology [11,12,33]. It is a highly complex task since it connects multiple different areas of computer vision. Three of them are motion or optical flow estimation, object understanding and understanding the 3D geometry of a scene. To show the large variety of existing motion segmentation approaches, we provide a brief overview presenting some fundamentally very different approaches tackling the same problem - approaches considering scene geometry, approaches focusing on general object segmentation (and thus rely more on appearance rather than scene geometry) and approaches that attempt to solve the problem of motion segmentation jointly, considering appearance as well as the geometrical structure. Most motion segmentation approaches rely on optical flow, except [32], which learn scene structure and motion coherently based on consecutive video frames rather than relying on point to point correspondences.

Geometrical based approaches relying on optical flow seek to find coherent motion patterns, while considering the scene geometry. These methods can be grouped into those that use projective geometry approximations [31,37] and those that use perspective projection [1,8,17,19]. [1,2,8] first attempt to simplify the observed motion field by compensating the flow field for camera rotation and then interpret the remaining flow using probabilistic models to segment the video into static environment and independently moving objects. This way they achieve highly accurate motion segmentations even videos that show complex scenes with high variation in depth.

Recently published approaches aim to learn motion patterns using neural networks. Those approaches directly take the optical flow as input, and motion segmentation patterns are learned without considering the scene geometry or the physical background behind the process of flow formation. Tokmakov et al.

learn motion patterns given optical flow [29,30]. A network segments a frame into static environment and moving objects given a flow represented as angle and magnitude separately. Besides their motion segmentation network, they benefit from the great ability of CNNs to learn object appearance. Jain et al. [9] learn motion patterns given the rgb-flow image as input. Despite the fact that both approaches do not take any geometrical information into account that can be extracted from optical flow, they achieve great performance on several standard video segmentation benchmarks [3,10,13,18,20,21,28].

In this work we present a new approach which, similar to [1,2,17], analyses the geometrical information provided by an optical flow image and reduces available information to its necessary minimum (the *flow angle* after compensation the optical flow for camera rotation) for the purpose of motion segmentation. Different from [1,2,17] a neural network is used for the final step of motion segmentation. Considering the geometrical background behind the process of flow formation allows us to generate an unlimited amount of synthetic training data, and thus the network can be trained in a self-supervised manner.

3 A Geometrical Analysis of the Flow Field

Optical flow describes where a pixel in the current frame will be in the next frame. These pixel displacements arise either due to the observer's motion or an object's motion. Object motion is very hard to predict, since objects move in many different ways. Their motion can be purely translational, rotational or both. Often their motion is articulated and thus can be described with neither rotation nor translation. In the following Section we review the formation process of optical flow due to observer's motion. The observer's motion or the camera motion can be translational as well as rotational in 3D. In this case the optical flow is determined by the camera motion itself (and speed), the camera's focal length and the scene depth. We first address the geometry of optical flow due to camera rotation only, which contains no information about independently moving objects or the scene depth. We continue with flow due to camera translation, which is informative in many regards, and thus is very valuable for the motion segmentation task.

3.1 Flow Due to Camera Rotation

Let f be the camera's focal length. A camera rotation is defined by its three rotational parameters (A, B, C). Given the three rotational parameters (A, B, C), we can compute the rotational optical flow vector at each pixel position (x, y) as follows [14]:[1]

$$\boldsymbol{v_r} = \begin{pmatrix} u_r \\ v_r \end{pmatrix} = \begin{pmatrix} \frac{A}{f}xy - Bf - \frac{B}{f}x^2 + Cy \\ Af + \frac{A}{f}y^2 - \frac{B}{f}xy - Cx \end{pmatrix} \tag{1}$$

[1] This equation only holds if rotation angles are small. However camera rotation is always independent of the scene depth regardless their amount.

The rotational flow vector $\boldsymbol{v_r}$ is independent of the scene depth, thus it can be simply subtracted from the optical flow $\boldsymbol{v}$ to "stabilize" the image.

3.2 Flow Due to Camera Translation

Let (U, V, W) be the translational motion of the camera relative to an object. Let (X, Y, Z) be the real world coordinates in 3D of a point that projects to (x, y) in the image. The motion field vector (u, v) at the image location (x, y) due to a translational motion is given by

$$\boldsymbol{v_t} = \begin{pmatrix} u_t \\ v_t \end{pmatrix} = \frac{1}{Z} \begin{pmatrix} -fU + xW \\ -fV + yW \end{pmatrix}. \tag{2}$$

The translational flow vector $\boldsymbol{v_t}$ is inversely proportional to the scene depth Z, thus a large flow magnitude might be due to high motion speed, or the pictured object is just very close to the camera. Just based on the flow magnitude, we are not able to distinguish between the two possible sources - speed and depth. The 2D translational motion direction at each point in the image is then given by the angle of the motion field vector (u, v) at image location (x, y):

$$\theta = \texttt{atan}(-fV + yW, -fU + xW) = \texttt{atan}(-V' + yW, -U' + xW). \tag{3}$$

The translational flow direction at a particular pixel (x, y) however is purely determined by the parameters (U', V', W). The focal length does not need to be known explicitly.

3.3 Flow Due to Camera Translation and Object Motion

Above we discussed how optical flow is formed due to camera translation. Now if we consider a moving object in the scene like a walking person, this will change the optical flow. In areas of the moving object this adds the object's motion to the optical flow. So we obtain

$$\boldsymbol{v} = \boldsymbol{v_t} + \boldsymbol{v_o}. \tag{4}$$

Objects might move at different speeds and in different directions. This changes the observed optical flow field. The optical flow's magnitude is now determined by three pixel-motion sources - the scene depth, the camera motion and the object motion. This makes drawing a conclusion based on the magnitude alone very hard. One can not distinguish what is actually moving in the world based on the flow's magnitude.

The flow's direction however is easy to interpret, since it is only determined by the observer's motion direction and the object's motion direction. Any deviation from the optical flow that is caused by the observer's translation v_t indicates some independent object motion in the scene. We use this information to generate a dataset to train a network for motion segmentation.

4 Learning Motion Patterns

Motion patterns in optical flow are often quite difficult to interpret directly. Camera rotation and translation couple the scene depth, which makes it impossible to judge whether an object is moving or not. Motion magnitude as well as direction are dependent on camera motion, object motion and depth, when the camera is rotating and translating simultaneously. Inspired by visual ecology and the purpose of human eye movements, we use a two step approach for motion segmentation (see Fig. 1). The two steps are as follows:

1. Compensate optical flow for rotation
   - Compensate the optical flow for the rotational component of the observer's motion, similar to the way that image stabilization is done on the human retina, which is done via small eye rotations. The rotation compensated flow is $\boldsymbol{v}$.
2. Segment optical flow $\boldsymbol{v}$ into static environment and moving objects
   - Given the flow $\boldsymbol{v}$ compute its direction θ at each pixel location.
   - A neural network MoA-Net (*Mo*tion *A*ngle - *Net*work) takes an angle image as input and generates per-pixel motion labels.

Rather than having the network learn complex geometrical dependencies, the fundamental idea is to break down the observed optical flow into a pattern that is easier to interpret. The input to the network - the *angle image* - is simpler and contains all of the motion information that can be obtained from optical flow.

In this work we assume the rotation to be known and present an approach that automatically segments the optical flow $\boldsymbol{v}$ into static environment and moving objects. We leave the step of estimating the camera's rotation and compensating the flow for rotation for future work.

4.1 Network Architecture

Our basic network architecture is adopted from [29,30]. Originally this network took as input the optical flow angle and flow magnitude - leading to a three dimensional input of size $[height \times width \times 2]$. Instead our network takes the angle image, which just has two dimensions $[height \times width]$, as input. The angles are in the range of $[-\pi, \ldots, \pi]$.

4.2 Training: Incorporating the Basics of Perspective Projection

Training a neural network for the task of motion segmentation usually requires large amounts of optical flow and its corresponding motion segmentations. The problem of using those datasets for training is that those datasets are often limited in size and the variety in scene geometry and motion is often restricted.

FlyingThings3D [16] is a relatively large synthetic flow dataset comprising 2700 videos, containing 10 stereo frames each. Along with these videos, ground truth optical flow, disparity, intrinsic, extrinsic camera parameters and object instance segmentation masks are provided. However this dataset doesn't picture realistic scenarios - random objects like tables, chairs and cars are flying in the 3D world.

Sintel [5,35] is a well-known optical flow dataset, containing 23 video sequences with 20 to 50 frames each. These short video sequences are taken from the computer animated movie by Blender. Thus the scenes are relatively realistic simulated. Videos come with ground truth optical flow, depth, intrinsic and extrinsic camera parameters and material segmentation.

Rather than relying on restricted datasets, the problem of motion segmentation is broken down into two small subproblems that can be each tackled separately. If a rotation compensated optical flow field is given, its geometry is easy to capture and motion information can be extracted. The optical flow's angle in areas of the static environment is completely determined by a translational motion direction of the camera that is projected onto the image plane. Moving objects move independently of the camera motion and thus are visible in the angle image due to its different motion direction. With this knowledge we can synthesize training data incorporating the physics of perspective projection for a motion segmentation network in an artificial manner.

Generating Training Data. For the purpose of motion segmentation we define a moving object as a *connected image region* that undergoes some *independent motion*. The connected image region can be of any size and shape - there are no limitations. True object motion can be quite complex, since objects can be deformable and articulated. If an object is articulated, each part might move independently of the other parts, e.g. a walking person. In case of a walking person, one arm might move forward while the other is standing still - here, although the body parts are physically connected, each part can move relatively independently of each other. The static environment undergoes a single pure translational motion due to the observers motion. Training data should contain these key criteria reflecting object motion and observer motion.
We generate training data for motion segmentation in 5 steps:

1. Generating connected object regions: To cover a large variety of different object shapes and sizes, we use the binary segmentations masks of FlyingThings3D [16,29] (Fig. 2a).
2. Modeling articulated object motion: To model object motion, each object region is split into n subregions using superpixels. n is a random number between one and ten. Splitting objects into subregions as shown in Fig. 2b leads to multiple different motion regions. In Fig. 2b we have eight motion regions including the region of static environment.
3. We assign to each motion region a translational 3D direction (Fig. 2c). A 3D translational direction is represented as a 3D unit vector. We generate a set of equally distributed translational motion direction on a sphere using the vertices of an icosahedron as approximation. Each vertex of an icosahedron represents a translational motion direction. To generate a large set of possible translational motion direction, we generate an icosahedron of frequency 50

which has 25002 vertices representing the set of translational motion directions.

4. Smoothing motion boundaries: To smooth motion boundaries within an object, we use a Gaussian filter with standard deviation $\sigma = 50$ (Fig. 2d). Object boundaries remain sharp.
5. We add random Gaussian noise with zero mean and standard deviation $\sigma = 0.1$ (Fig. 2e).

This procedure to generate training data is entirely independent of any color images or other labeled training data. It incorporates all geometrical information required to segment independently moving objects. This abstraction - reducing objects to *connected image regions* that undergo *independent motion* - allows us to train a network with unlimited training data in a fully unsupervised manner.

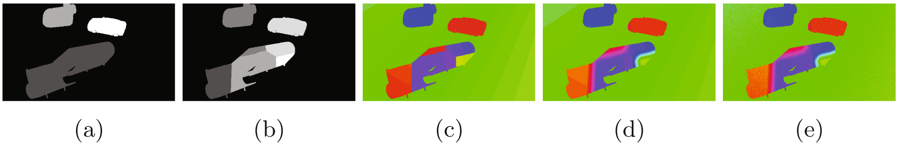

Fig. 2. Generating Training Data. The abstract object definition reduces an object to *connected image regions* that undergo *independent motion.* (a)–(e) show the process of generating abstract objects for the motion segmentation task.

5 Experiments and Results

We evaluate our work on Sintel [5,35] and FlyingThings3D [16]. We generated additional motion segmentation ground truth for Sintel to use this dataset for evaluation. Both datasets provide camera motion information, which allows us to evaluate the performance of MoA-Net, which requires flow angles of the rotation compensated flow field as input. We compare our work to two recently published motion segmentations approaches [9,30]. Both approaches are learning based approaches that attempt to learn motion patterns given the optical flow as input. In combination with a neural network that produces object segmentations based on appearance, both approaches have shown great results on a variety of different datasets [3,10,13,18,20,21,28]. For comparison purposes, we extract the motion segmentation network of both works and compare their performance on ground truth optical flow with our proposed method. The very modular motion segmentation pipeline of Tokmakov et al. [30] as well as of Jain et al. [9] allows us to analyze their "motion-stream" exclusively.

Compensating for Camera Rotation. Besides ground truth optical flow, Sintel and FlyingThings3D provide ground truth extrinsic and intrinsic camera matrices. This allows us to compensate the flow for camera rotation. We move image coordinates x_t along the optical flow and obtain new image coordinates x_{t+1}. The new image coordinates x_{t+1} are transformed into 3D camera coordinates X_{t+1}. We compute the camera motion (rotation and translation) between two consecutive frames and undo the camera rotation in 3D. The new camera coordinates X_{trans} (after undoing the camera's rotation) are projected back onto the 2D image plane. The rotation compensated flow can be obtained from the pixel displacement between image coordinates x_t and x_{trans}.

Evaluation. We use the evaluation scheme of [20]. We show results on two different motion segmentation networks and compare their performance with our motion network on Sintel and the test set of FlyingThings3D (Figs. 3 and 4).
Jain et al. train a motion segmentation network given rgb-flow images as input. For training, they used estimated optical flow images in rgb-format. Since no motion segmentation are available for ImageNet [23], they propose a procedure to produce (pseudo)-ground truth segmentations based on the provided object bounding boxes, the segmentations of their appearance network and the appearance of the estimated optical flow. Flow images are discarded from the training set, if average rgb-flow inside an object bounding box differs not sufficiently from the background's optical flow. Their segmentations are rather conservative - they often segment just a small portion of the moving object or nothing, which leads to an overall low performance of their motion segmentation network. On both datasets - Sintel and FlyingThings3D - their performance is rather low. One might argue that moving objects in Sintel and FlyingThings3D are quite different from objects that the network trained on ImageNet has seen before. Also, their automatic procedure to generate (pseudo)-ground truth significantly limits the variability of motion fields.
Tokmakov et al. trained their network on ground truth optical flow provided by the FlyingThings3D dataset. Each flow vector is represented using polar coordinates (flow magnitude and angle) during training. On Sintel as well as FlyingThings3D they show overall a good performance. If a video scene shows high variance in depth as in the bamboo video sequences of Sintel (Figs. 5 and 6), their segmentation is highly depth dependent, which leads to erroneous motion segmentations. Especially in those cases, MoA-Net outperforms both other motion segmentation networks by a large margin.
MoA-Net (ours) is trained purely on translational angle fields. This allows for producing motion segmentations that are completely independent upon the scene depth.

Results. On Sintel we outperform Tokmakov et al. by 4% points using the J-Mean metric and by more than 7% points regarding the F-Mean (see Table 1). On FlyingThings3D, the motion segmentation network of Tokmakov et al. produces high quality motion segmentation masks. Their accuracy in terms of IoU differs from their performance on Sintel by a large margin (39% points). This significant difference is very likely due to the similar nature of training and test data.

Table 1. Comparison to state-of-the-art. We compare our motion segmentation network with two recent motion segmentation networks that segment optical flow into static background and independently moving objects. The top results are highlighted in **blue**.

Motion Segmentation: Sintel						
Motion						
	J Mean	J Recall	J Decay	F Mean	F Recall	F Decay
	↑	↑	↓	↑	↑	↓
Tokmakov et al. [29,30]	50.38	**55.43**	45.32	52.43	54.95	45.58
Jain et al. [9]	30.27	24.78	32.72	28.07	14.02	31.89
Ours	**55.13**	55.24	**26.62**	**59.94**	**61.67**	**16.76**

When our MoA-Net is trained on the same ground truth flow as Tokmakov et al., but using only the optical flow's angle after compensating for camera rotation, we outperform their method (91.12% versus 89.13% - see Table 2)). Our proposed motion segmentation network, however, is trained in a self-supervised manner. We show significantly better performance than Jain et al. on Sintel as well as FlyingThings3D. We achieve state-of-the-art results on Sintel, whereas on FlyingThings3D we rank second best after Tokmakov et al.

Tokmakov et al. and Jain et al. do not need any preprocessing of the optical flow, however, here we show that a more analytical approach, which includes a step of preprocessing the optical flow - compensating for camera rotation, has a high potential for further improvements and solving the task of motion segmentation without the need of large training datasets.

Table 2. Comparison of motion networks trained on different training data and tested on FlyingThings3D-Test. Tokmakov et al. and ours-FT3D are trained using the provided ground truth optical flow of FlyingThings3D, Jain et al. relies on estimated optical flow of a subset of videos from ImageNet, and ours is trained on fully automatically generated training data as described in Sect. 4.2.

Motion Segmentation: FlyingThings3D-Test						
Motion						
	J Mean	J Recall	J Decay	F Mean	F Recall	F Decay
	↑	↑	↓	↑	↑	↓
Tokmakov et al. [29,30]	89.13	98.40	**−2.11**	93.55	98.54	**−2.29**
Jain et al. [9]	21.57	6.47	2.51	30.04	8.77	1.85
Ours (flow angle FT3D)	**91.12**	**99.78**	−0.02	**94.33**	**99.63**	−0.41
Ours (self-supervised)	75.53	95.76	3.55	82.25	97.65	1.68

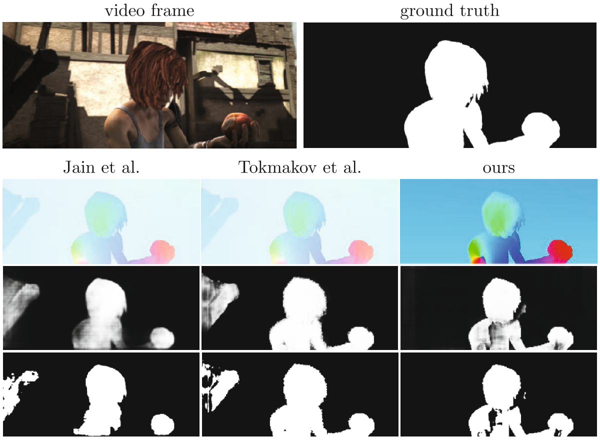

Fig. 3. Sintel - alley1: *first row*: input frame and ground truth motion segmentation. *Second row*: input to the motion segmentation network of the two different methods used for comparison an our input - optical flow as rgb image, optical flow in its angle and magnitude representation, angle of the rotation compensated flow. *Third row*: raw motion network output for each method. *Fourth row*: motion segmentation of each method (Color figure online)

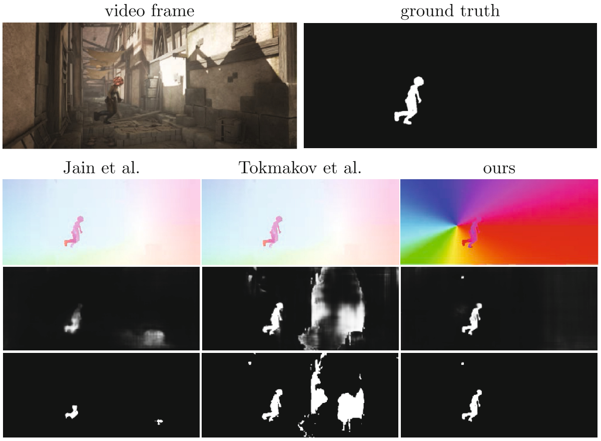

Fig. 4. Sintel - alley2: *first row*: input frame and ground truth motion segmentation. *Second row*: input to the motion segmentation network of the two different methods used for comparison an our input - optical flow as rgb image, optical flow in its angle and magnitude representation, angle of the rotation compensated flow. *Third row*: raw motion network output for each method. *Fourth row*: motion segmentation of each method (Color figure online)

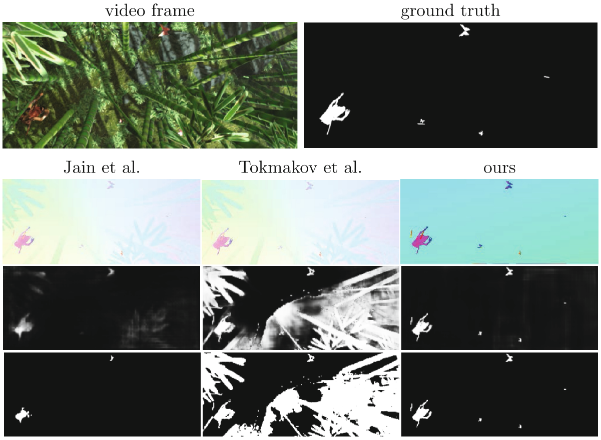

Fig. 5. Sintel - bamboo1: *first row*: input frame and ground truth motion segmentation. *Second row*: input to the motion segmentation network of the two different methods used for comparison an our input - optical flow as rgb image, optical flow in its angle and magnitude representation, angle of the rotation compensated flow. *Third row*: raw motion network output for each method. *Fourth row*: motion segmentation of each method (Color figure online)

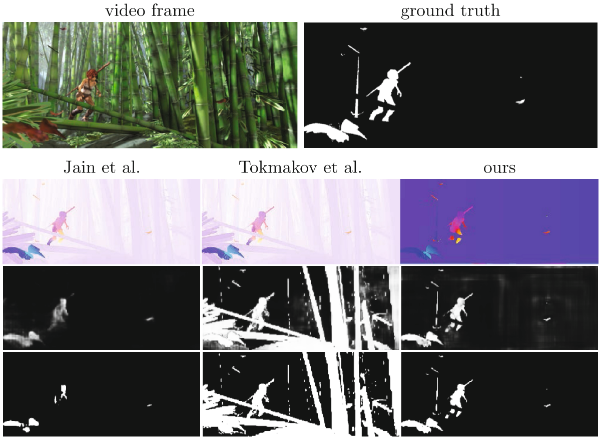

Fig. 6. Sintel - bamboo2: *first row*: input frame and ground truth motion segmentation. *Second row*: input to the motion segmentation network of the two different methods used for comparison an our input - optical flow as rgb image, optical flow in its angle and magnitude representation, angle of the rotation compensated flow. *Third row*: raw motion network output for each method. *Fourth row*: motion segmentation of each method (Color figure online)

6 Conclusion

We present a new approach for learning of motion segmentation in a self-supervised way. We break the problem of motion segmentation down into two smaller subtasks: (1) compensating the flow for camera rotation and (2) segmenting the remaining flow angle field into static environment and moving objects. This has led to an "abstract" definition of an moving object, which allowed us to synthesise training data in a fully automatic way and makes it possible to use a CNN for for the task of motion segmentation while simultaneously ensuring a correct interpretation of the scenes geometry. We show significant improvement in performance for motion segmentation among other motion segmentation networks, as shown in our experiments.

However, one has to note that the first step, which is compensating the flow for camera rotation, still remains subject of current research [1,2]. In future work we will investigate more the task of compensating the flow for camera rotation, which is comparable to motion compensation on our retina that is done by small eye rotations.

References

1. Bideau, Pia, Learned-Miller, Erik: It's moving! A probabilistic model for causal motion segmentation in moving camera videos. In: Leibe, Bastian, Matas, Jiri, Sebe, Nicu, Welling, Max (eds.) ECCV 2016. LNCS, vol. 9912, pp. 433–449. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_26
2. Bideau, P., RoyChowdhury, A., Menon, R.R., Learned-Miller, E.: The best of both worlds: combining CNNs and geometric constraints for hierarchical motion segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 508–517 (2018)
3. Brox, Thomas, Malik, Jitendra: Object segmentation by long term analysis of point trajectories. In: Daniilidis, Kostas, Maragos, Petros, Paragios, Nikos (eds.) ECCV 2010. LNCS, vol. 6315, pp. 282–295. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15555-0_21
4. Brox, T., Malik, J.: Large displacement optical flow: descriptor matching in variational motion estimation. IEEE Trans. Pattern Anal. Mach. Intell. **33**(3), 500–513 (2011)
5. Butler, Daniel J., Wulff, Jonas, Stanley, Garrett B., Black, Michael J.: A naturalistic open source movie for optical flow evaluation. In: Fitzgibbon, Andrew, Lazebnik, Svetlana, Perona, Pietro, Sato, Yoichi, Schmid, Cordelia (eds.) ECCV 2012. LNCS, vol. 7577, pp. 611–625. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33783-3_44
6. Horn, B.K., Schunck, B.G.: Determining optical flow. Artif. Intell. **17**(1–3), 185–203 (1981)
7. Hur, Junhwa, Roth, Stefan: Joint optical flow and temporally consistent semantic segmentation. In: Hua, Gang, Jégou, Hervé (eds.) ECCV 2016. LNCS, vol. 9913, pp. 163–177. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46604-0_12
8. Irani, M., Anandan, P.: A unified approach to moving object detection in 2D and 3D scenes. **20**(6), 577–589 (1998)

9. Jain, S., Xiong, B., Grauman, K.: Fusionseg: learning to combine motion and appearance for fully automatic segmention of generic objects in videos. In: CVPR (2017)
10. Jain, Suyog Dutt, Grauman, Kristen: Supervoxel-consistent foreground propagation in video. In: Fleet, David, Pajdla, Tomas, Schiele, Bernt, Tuytelaars, Tinne (eds.) ECCV 2014. LNCS, vol. 8692, pp. 656–671. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_43
11. Land, M.F.: Motion and vision: why animals move their eyes. J. Comp. Physiol. A **185**(4), 341–352 (1999)
12. Lappe, M., Hoffmann, K.P., et al.: Optic flow and eye movements. Int. Rev. Neurobiol. 29–50 (2000)
13. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2192–2199 (2013)
14. Longuet-Higgins, H.C., Prazdny, K., et al.: The interpretation of a moving retinal image. Proc. R. Soc. Lond. B **208**(1173), 385–397 (1980)
15. Lucas, B.D., Kanade, T., et al.: An iterative image registration technique with an application to stereo vision (1981)
16. Mayer, N., et al.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4040–4048 (2016)
17. Narayana, M., Hanson, A., Learned-Miller, E.: Coherent motion segmentation in moving camera videos using optical flow orientations, pp. 1577–1584 (2013)
18. Ochs, P., Malik, J., Brox, T.: Segmentation of moving objects by long term video analysis. IEEE Trans. Pattern Anal. Mach. Intell. **36**(6), 1187–1200 (2014)
19. Ogale, A.S., Fermüller, C., Aloimonos, Y.: Motion segmentation using occlusions **27**(6), 988–992 (2005)
20. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Computer Vision and Pattern Recognition (2016)
21. Prest, A., Leistner, C., Civera, J., Schmid, C., Ferrari, V.: Learning object class detectors from weakly annotated video. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3282–3289. IEEE (2012)
22. Revaud, J., Weinzaepfel, P., Harchaoui, Z., Schmid, C.: EpicFlow: edge-preserving interpolation of correspondences for optical flow. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1164–1172 (2015)
23. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
24. Sawhney, H.S., Guo, Y., Kumar, R.: Independent motion detection in 3d scenes. IEEE Trans. Pattern Anal. Mach. Intell. **22**(10), 1191–1199 (2000)
25. Sevilla-Lara, L., Sun, D., Jampani, V., Black, M.J.: Optical flow with semantic segmentation and localized layers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3889–3898 (2016)
26. Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2432–2439. IEEE (2010)
27. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8934–8943 (2018)

28. Tang, K., Sukthankar, R., Yagnik, J., Fei-Fei, L.: Discriminative segment annotation in weakly labeled video. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2483–2490 (2013)
29. Tokmakov, P., Alahari, K., Schmid, C.: Learning motion patterns in videos. In: CVPR (2017)
30. Tokmakov, P., Alahari, K., Schmid, C.: Learning video object segmentation with visual memory. In: ICCV (2017)
31. Torr, P.H.: Geometric motion segmentation and model selection. Philos. Trans. R. Soc. Lond. A: Math. Phys. Eng. Sci. **356**(1740), 1321–1340 (1998)
32. Vijayanarasimhan, S., Ricco, S., Schmid, C., Sukthankar, R., Fragkiadaki, K.: SfM-Net: learning of structure and motion from video. arXiv preprint arXiv:1704.07804 (2017)
33. Walls, G.: The evolutionary history of eye movements. Vis. Res. **2**(1–4), 69–80 (1962)
34. Wang, J.Y., Adelson, E.H.: Representing moving images with layers. IEEE Trans. Image Process. **3**(5), 625–638 (1994)
35. Wulff, Jonas, Butler, Daniel J., Stanley, Garrett B., Black, Michael J.: Lessons and insights from creating a synthetic optical flow benchmark. In: Fusiello, Andrea, Murino, Vittorio, Cucchiara, Rita (eds.) ECCV 2012. LNCS, vol. 7584, pp. 168–177. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33868-7_17
36. Wulff, J., Sevilla-Lara, L., Black, M.J.: Optical flow in mostly rigid scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
37. Zamalieva, D., Yilmaz, A.: Background subtraction for the moving camera: a geometric approach **127**, 73–85 (2014)