It's Moving! A Probabilistic Model for Causal Motion Segmentation in Moving Camera Videos

Pia Bideau, Erik Learned-Miller

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

Abstract. The human ability to detect and segment moving objects works in the presence of multiple objects, complex background geometry, motion of the observer, and even camouflage. In addition to all of this, the ability to detect motion is nearly instantaneous. While there has been much recent progress in motion segmentation, it still appears we are far from human capabilities. In this work, we derive from first principles a likelihood function for assessing the probability of an optical flow vector given the 2D motion direction of an object. This likelihood uses a novel combination of the angle and magnitude of the optical flow to maximize the information about how objects are moving differently. Using this new likelihood and several innovations in initialization, we develop a motion segmentation algorithm that beats current state-of-the-art methods by a large margin. We compare to five state-of-the-art methods on two established benchmarks, and a third new data set of camouflaged animals, which we introduce to push motion segmentation to the next level.

1 Introduction

"Motion is a powerful cue for image and scene segmentation in the human visual system. This is evidenced by the ease with which we see otherwise perfectly camouflaged creatures as soon as they move." –Philip Torr [1]

How can we match the ease and speed with which humans and other animals detect motion? This remarkable capability works in the presence of complex background geometry, camouflage, and motion of the observer. Figure 1 shows a frame from a video of a "walking stick" insect. Despite the motion of the camera, the rarity of the object, and the high complexity of the background geometry, the insect is immediately visible as soon as it starts moving.

To develop such a motion segmentation system, we re-examined classical methods based upon perspective projection, and developed a new probabilistic model which accurately captures the information about 3D motion in each observed optical flow vector $\boldsymbol{v}$. First, we estimate the portion of the optical flow due to rotation, and subtract it from $\boldsymbol{v}$ to produce $\boldsymbol{v_t}$, the translational portion of the optical flow. Next, we derive a new *conditional flow angle likelihood*

Fig. 1: **Where is the camouflaged insect?** Before looking at Figure 2, which shows the ground truth localization of this insect, try identifying the insect. While it is virtually impossible to see without motion, it immediately "pops out" to human observers as it moves in the video (see supplementary material).

$\mathcal{L} = p(\theta_{\boldsymbol{v_t}} \mid M, \|\boldsymbol{v_t}\|)$, the probability of observing a particular flow angle $\theta_{\boldsymbol{v_t}}$ given a model M of the angle part of a particular object's (or the background's) motion field and the flow magnitude $\|\boldsymbol{v_t}\|$.

M, which we call an *angle field*, describes the motion *directions* of an object in the image plane. It is a function of the object's relative motion (U, V, W) and the camera's focal length f, but can be computed more directly from a set of *motion field parameters* $(U', V', W) = (fU, fV, W)_2$, where the "2" subscript indicates L_2 normalization.

Our new angle likelihood helps us to address a fundamental difficulty of motion segmentation: the ambiguity of 3D motion given a set of noisy flow vectors. While we cannot eliminate this problem, the angle likelihood allows us to weigh the evidence for each image motion properly based on the optical flow. In particular, when the underlying image motion is very small, moderate errors in the optical flow can completely change the apparent motion direction (i.e., the angle of the optical flow vector). When the underlying image motion is large, typical errors in the optical flow will not have a large effect on apparent motion direction. This leads to the critical observation that small optical flow vectors are less informative about motion than large ones. Our derivation of the angle likelihood (Section 3) quantifies this notion and makes it precise in the context of a Bayesian model of motion segmentation.

We evaluate our method on three diverse data sets, achieving state-of-the-art performance on all three. The first is the widely used Berkeley Motion Segmentation (BMS-26) database [2, 3], featuring videos of cars, pedestrians, and other common scenes. The second is the Complex Background Data Set [4], designed to test algorithms' abilities to handle scenes with highly variable depth. Third, we introduce a new and even more challenging benchmark for motion segmentation algorithms: the *Camouflaged Animal Data Set*. The nine (moving camera) videos in this benchmark exhibit camouflaged animals that are difficult to see in a single frame, but can be detected based upon their motion across frames.

2 Related Work

A large number of motion segmentation approaches have been proposed, including [2, 4–25]. We focus our review on recent methods.

Many methods for motion segmentation work by tracking points or regions through multiple frames to form motion trajectories, and grouping these trajectories into coherent moving objects [2, 17, 18, 20, 26]. Elhamifar and Vidal [26]

Fig. 2: **Answer:** the insect from Figure 1 in shown in red. The insect is trivial to see in the original video, though extremely difficult to identify in a still image. In addition to superior results on standard databases, our method is also one of the few that can detect objects is such complex scenes.

track points through multiple images and show that rigid objects are represented by low-dimensional subspaces in the space of tracks. They use sparse subspace clustering to identify separate objects. Brox and Malik [2] define a pairwise metric on multi-frame trajectories so that they may be clustered to perform motion segmentation. Fragkiadaki et al. [20] detect discontinuities of the embedding density between spatially neighboring trajectories. These discontinuities are used to infer object boundaries and perform segmentation. Papazoglou and Ferrari [17] develop a method that looks both forward and backward in time, using flow angle and flow magnitude discontinuities, appearance modeling, and superpixel mapping across images to connect independently moving objects across frames. Keuper et al. [18] also track points across multiple frames and use minimum cost multicuts to group the trajectories.

These trajectory-based methods are *non-causal*. To segment earlier frames, the knowledge of future frames is necessary. We propose a causal method, relying only on the flow between two frames and information passed forward from previous frames. Despite this, we outperform trajectory-based methods, which in general tend to a depth dependent motion segmentation (see Experiments).

Another set of methods analyze optical flow between a pair of frames, grouping pixels into regions whose flow is consistent with various motion models. Torr [1] develops a sophisticated probabilistic model of optical flow, building a mixture model that explains an arbitrary number of rigid components within the scene. Interestingly, he assigns different types of motion models to each object based on model fitting criteria. His approach is fundamentally based on projective geometry rather based directly on perspective projection equations, as in our approach. Horn has identified drawbacks of using projective geometry in such estimation problems and has argued that methods based directly on perspective projection are less prone to overfitting in the presence of noise [27]. Zamalieva et al. [16] present a combination of methods that rely on homographies and fundamental matrix estimation. The two methods have complementary strengths, and the authors attempt to select among the best dynamically. An advantage of our method is that we do not depend upon the geometry of the scene to be well-approximated by a group of homographies, which enables us to address videos with very complex background geometries. Narayana et al. [4] remark that for translational only motions, the angle field of the optical flow will consist of one of a set of canonical angle fields, one for each possible motion direction, regardless of the focal length. They use these canonical angle fields as a basis with which to

segment a motion image. However, they do not handle camera rotation, which is a significant limitation.

Another set of methods using occlusion events in video to reason about depth ordering and independent object motion [19, 28]. Ogale et al. [28] use occlusion cues to further disambiguate non-separable solutions to the motion segmentation problem. Taylor et al. [19] introduce a causal framework for integrating occlusion cues by exploiting temporary consistency priors to partition videos into depth layers.

3 Methods

The *motion field* of a scene is a 2D representation of 3D motion. Motion vectors, describing the displacement in 3D, are projected onto the image plane forming a 2D motion field. This field is created by the movement of the camera relative to a stationary environment and the additional motion of independently moving objects. We use the optical flow, or estimated motion field, to segment each video image into static environment and independently moving objects.

The observed flow field consists of the flow vectors $\boldsymbol{v}$ at each pixel in the image. Let $\boldsymbol{m}$ be the flow vectors describing the motion field caused only by a rotating and translating camera in its stationary 3D environment – $\boldsymbol{m}$ does not include motion of other independently moving objects. The flow vectors $\boldsymbol{m}$ can be decomposed in a *translational component* $\boldsymbol{m_t}$ and a *rotational component* $\boldsymbol{m_r}$. Let the direction or angle of a flow vector of a translational camera motion at a particular pixel (x, y) be $\theta_{\boldsymbol{m_t}}$.

When the camera is only translating, there are strong constraints on the optical flow field – the *direction* $\theta_{\boldsymbol{m_t}}$ of the motion at each pixel is determined by the camera translation (U, V, W), the image location of the pixel (x, y), and the camera's focal length f, and has no dependence on scene depth [29].

$$\theta_{\boldsymbol{m_t}} = \arctan(W \cdot y - V \cdot f, W \cdot x - U \cdot f) \tag{1}$$

$$= \arctan(W \cdot y - V', W \cdot x - U') \tag{2}$$

The collection of $\theta_{\boldsymbol{m_t}}$ forms a translational angle field M representing the camera's translation direction on the 2D image plane.

Simultaneous camera rotation and translation, however, couple the scene depth and the optical flow, making it much harder to assign pixels to the right angle field M described by the estimated translation parameters (U', V', W).

To address this, we wish to subtract off the flow vectors $\boldsymbol{m_r}$ describing the rotational camera motion field from the observed flow vectors $\boldsymbol{v}$ to produce a flow $\boldsymbol{v_t}$ comprising camera translation only. The subsequent assignment of flow vectors to particular angle fields is thus greatly simplified. However estimating camera rotation in the presence of multiple motions is challenging. We organize the Methods section as follows:

In Section 3.1, we describe how all frames after the first frame are segmented, using the segmentation from the previous frame and our novel angle likelihood.

After reviewing Bruss and Horn's motion estimation technique [30] in Section 3.2, Section 3.3 describes how our method is initialized in the first frame, including a novel process for estimating camera motion in the presence of multiple motions.

3.1 A probabilistic model for motion segmentation

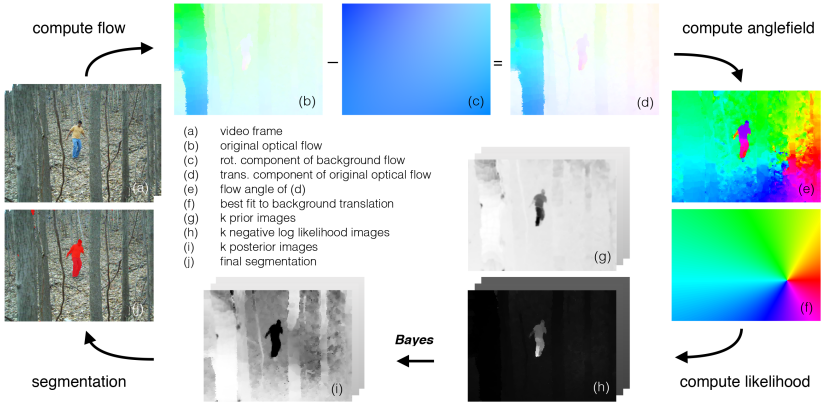

Fig. 3: **Our segmentation procedure.** Given the optical flow (b) the camera rotation is estimated. Then, the flow $\boldsymbol{m_r}$ due to camera rotation defined by the motion parameters (A, B, C) (c) is subtracted from the optical flow $\boldsymbol{v}$ to produce a translational flow $\boldsymbol{v_t}$. The flow angles $\theta_{\boldsymbol{v_t}}$ of $\boldsymbol{v_t}$ are shown in (e). The best fitting translation parameters (U', V', W) to the static environment of $\boldsymbol{v_t}$ yield an estimated angle field M (f), which clearly shows the forward motion of the camera (rainbow focus of expansion pattern) not visible in the original angle field. The motion component priors (g) and negative log likelihoods (h) yield the posteriors (i) and the final segmentation (j).

Given a prior motion segmentation of frame $t-1$ into k different motion components and an optical flow from frames t and $t+1$, segmenting frame t requires several ingredients: **a)** the *prior* probabilities $p(M_j)$ for each pixel that it is assigned to a particular angle field M_j, **b)** the estimate of the translational angle field M_j, $1 \leq j \leq k$ to be able to model the motion for each of the k motion components from the previous frame, **c)** for each pixel position, a *likelihood* $\mathcal{L}_j = p(\boldsymbol{v_t} \mid M_j)$, the probability of observing a flow vector $\boldsymbol{v_t}$ under an estimated angle field M_j, and **d)** the prior probability $p(M_{k+1})$ and angle likelihoods $\mathcal{L}_{k+1}$ given an angle field M_{k+1} to model a *new motion*. Given these priors and likelihoods, we use Bayes' rule to obtain a *posterior* probability for each translational angle field at each pixel location. We have

$$p(M_j \mid \boldsymbol{v_t}) \propto p(\boldsymbol{v_t} \mid M_j) \cdot p(M_j) \tag{3}$$

We directly use this posterior for segmentation. We now describe how the above quantities are computed.

Propagating the posterior for a new prior. We start from the optical flow of Sun et al. [31] (Figure 3b). We then create a prior at each pixel for each angle field M_j in the new frame (Figure 3g) by propagating the posterior from the previous frame (Figure 3i) in three steps.

1. Use the previous frame's flow to map posteriors from frame $t-1$ (Figure 3i) to new positions in frame t.
2. Smooth the mapped posterior in the new frame by convolving with a spatial Gaussian, as done in [4,32]. This implements the idea that object locations in future frames are likely to be close to their locations in previous frames.
3. Renormalize the smoothed posterior from the previous frame to form a proper probability distribution at each pixel location, which acts as the prior on the k motion components for the new frame (Figure 3g). Finally, we set aside a probability of $1/(k+1)$ for the prior of a new motion component, while rescaling the priors for the pre-existing motions to sum to $k/(k+1)$.

Estimating and removing rotational flow. We use the prior for the motion component of the static environment to weight pixels for estimating the current frame's flow due to the camera motion. We estimate the camera translation parameters (U', V', W) and rotation parameters (A, B, C) using a modified version of the Bruss and Horn algorithm [30] (Section 3.2). As described above, we then render the flow angle independent of the unknown scene depth by subtracting the estimated rotational flow (Figure 3c) from the original flow (Figure 3b) to produce an estimate of the flow without influences of camera rotation (Fig. 3d). For each flow vector we compute:

$$\hat{\boldsymbol{v}}_{\boldsymbol{t}} = \boldsymbol{v} - \hat{\boldsymbol{m}}_{\boldsymbol{r}}(\hat{A}, \hat{B}, \hat{C}) \tag{4}$$

$$\theta_{\boldsymbol{v_t}} = \measuredangle(\hat{\boldsymbol{v}}_{\boldsymbol{t}}, \boldsymbol{n}) \tag{5}$$

Where $\boldsymbol{n}$ is a unit vector $[1, 0]^T$.

For each additional motion component j besides the static environment, we estimate 3D translation parameters (U', V', W) using the segment priors to select pixels, weighted according to the prior, such that the motion perceived from video frame t to $t+1$ is described by j independent angle fields M_j.

The flow angle likelihood. Once we have obtained a translational flow field by removing the rotational flow, we use each flow vector $\boldsymbol{v_t}$ to decide which motion component it belongs to. Most of the information about the 3D motion direction is contained in the flow angle, not the flow magnitude. This is because for a given translational 3D motion direction (relative to the camera), the flow angle is completely determined by that motion and the location in the image, whereas the flow magnitude is a function of the object's depth, which is unknown. However, as discussed above, the *amount of information* in the flow angle depends upon the flow magnitude–flow vectors with greater magnitude are much more reliable indicators of true motion direction. This is why it is critical to formulate the angle likelihood conditioned on the flow magnitude.

Other authors have used flow angles in motion segmentation. For example, Papazoglou and Ferrari [17] use both a gradient of the optical flow and a separate function of the flow angle to define motion boundaries. Narayana et al. [4]

use *only* the optical flow angle to evaluate motions. But our derivation gives a principled and effective method of using the flow angle and magnitude together to mine accurate information from the optical flow. In particular, we show that while (under certain mild assumptions) the translational magnitudes alone have no information about which motion is most likely, the magnitudes play an important role in specifying the *informativeness* of the flow angles. In our experiments section, we demonstrate that failing to condition on flow magnitudes in this way results in greatly reduced performance over our derived model.

We now derive the key element of our method, the *conditional flow angle likelihood* $p(\theta_{\boldsymbol{v_t}} \mid M_j, \|\boldsymbol{v_t}\|)$, the probability of observing a flow direction $\theta_{\boldsymbol{v_t}}$ given that a pixel was part of a motion component undergoing the 2D motion direction M_j, and that the flow magnitude was $\|\boldsymbol{v_t}\|$. We make the following modeling assumptions:

1. We assume the observed translational flow $\boldsymbol{v_t} = (\|\boldsymbol{v_t}\|, \theta_{\boldsymbol{v_t}})$ at a pixel is a noisy observation of the translational motion field $\boldsymbol{m_t} = (\|\boldsymbol{m_t}\|, \theta_{\boldsymbol{m_t}})$:

$$\boldsymbol{v_t} = \boldsymbol{m_t} + \eta, \tag{6}$$

   where η is independent 2D Gaussian noise with zero mean and circular but unknown covariance.
2. We assume the translational motion field magnitude $\|\boldsymbol{m_t}\|$ is statistically independent of the translation angle field M created by the estimated 3D translation parameters (U', V', W). It follows that $\|\boldsymbol{v_t}\| = \|\boldsymbol{m_t}\| + \eta$ is also independent of M, and hence $p(\|\boldsymbol{v_t}\| \mid M) = p(\|\boldsymbol{v_t}\|)$.

With these assumptions, we have

$$\begin{aligned} p(\boldsymbol{v_t} \mid M_j) &\overset{(1)}{=} p(\|\boldsymbol{v_t}\|, \theta_{\boldsymbol{v_t}} \mid M_j) && (7)\\ &= p(\theta_{\boldsymbol{v_t}} \mid \|\boldsymbol{v_t}\|, M_j) \cdot p(\|\boldsymbol{v_t}\| \mid M_j) && (8)\\ &\overset{(2)}{=} p(\theta_{\boldsymbol{v_t}} \mid \|\boldsymbol{v_t}\|, M_j) \cdot p(\|\boldsymbol{v_t}\|) && (9)\\ &\propto p(\theta_{\boldsymbol{v_t}} \mid \|\boldsymbol{v_t}\|, M_j), && (10) \end{aligned}$$

where the numbers over each equality give the assumption that is invoked. Equation (10) follows since $p(\|\boldsymbol{v_t}\|)$ is constant across all estimated angle fields.

We model $p(\theta_{\boldsymbol{v_t}} \mid \|\boldsymbol{v_t}\|, M)$ using a *von Mises* distribution $\mathcal{V}(\mu, \kappa)$ with parameters μ, the preferred direction, and concentration parameter κ. We set $\mu = \theta_{\boldsymbol{m_t}}$, since $\theta_{\boldsymbol{m_t}}$ is the most likely direction assuming a noisy observation of a translational motion θ_{v_t}. To set κ, we observe that when the ground truth flow magnitude $\|\boldsymbol{m_t}\|$ is small, the distribution of observed angles θ_{v_t} will be near uniform (see Figure 4, $\boldsymbol{m_t} = (0, 0)$), whereas when $\|\boldsymbol{m_t}\|$ is large, the observed angle $\theta_{\boldsymbol{v_t}}$ is likely to be close to the flow angle $\theta_{\boldsymbol{m_t}}$ (Figure 4, $\boldsymbol{m_t} = (2, 0)$). We can achieve this basic relationship by setting $\kappa = a(\|\boldsymbol{m_t}\|)^b$, where a and b are parameters that give added flexibility to the model. Since we don't have direct access to $\|\boldsymbol{m_t}\|$, we use $\|\boldsymbol{v_t}\|$ as a surrogate, yielding

$$p(\theta_{\boldsymbol{v_t}} \mid \|\boldsymbol{v_t}\|, M_j) \propto \mathcal{V}(\theta_{\boldsymbol{v_t}}; \mu = \theta_{\boldsymbol{m_t}}, \kappa = a\|\boldsymbol{v_t}\|^b). \tag{11}$$

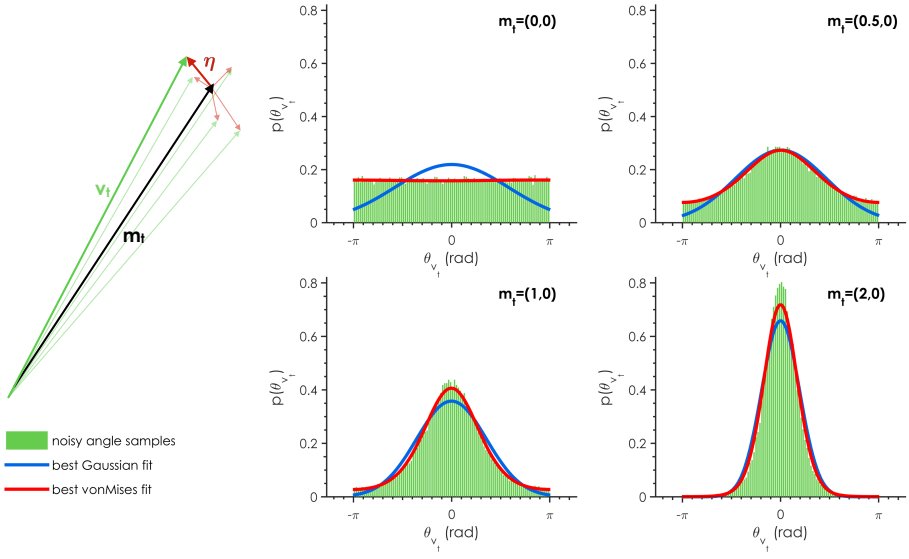

Fig. 4: **The von Mises distribution.** When a motion field vector $\boldsymbol{m_t}$ is perturbed by added Gaussian noise η (figure top left), the resulting distribution over optical flow angles θ_{v_t} is well-modeled by a *von Mises* distribution. The figure shows how small motion field vectors result in a broad distribution of angles after noise is added, while larger magnitude motion field vectors result in a narrower distribution of angles. The red curve shows the best von Mises fit to these sample distributions and the blue curve shows the lower quality of the best Gaussian fit.

Note that this likelihood treats zero-length translation vectors as uninformative– it assigns them the same likelihood under all motions. This makes sense, since the direction of a zero-length optical flow vector is essentially random. Similarly, the longer the optical flow vector, the more reliable and informative it becomes.

Likelihood of a new motion. Lastly, with no prior information about new motions, we set $p(\theta_{\boldsymbol{v_t}} \mid \|\boldsymbol{v_t}\|, M_j) = \frac{1}{2\pi}$, a uniform distribution.

Once we have priors and likelihoods, we compute the posteriors (Equation 3) and label each pixel as

$$L = \arg\max_j p(M_j \mid \boldsymbol{v_t}). \tag{12}$$

3.2 Bruss and Horn's motion estimation.

To estimate the translation parameters (U', V', W) of the camera relative to the static environment, we use the method of Bruss and Horn [30] and apply it to pixels selected by the prior of M_j. The observed optical flow vector $\boldsymbol{v}_i$ at pixel i can be decomposed as $\boldsymbol{v}_i = \boldsymbol{p}_i + \boldsymbol{e}_i$, where $\boldsymbol{p}_i$ is the component of $\boldsymbol{v}_i$ in the predicted direction θ_{m_t} and $\boldsymbol{e}_i$ is the component orthogonal to $\boldsymbol{p}_i$. The authors find the motion parameters that minimizes the sum of these "error" components

$\boldsymbol{e}_i$. The optimization for translation-only is

$$\underset{U',V',W}{\arg\min} \sum_i \|\boldsymbol{e}_i(\boldsymbol{v}_i, U', V', W)\|, \tag{13}$$

where $(U', V', W) = (Uf, Vf, W)$ are the three translation parameters. Since we do not know the focal length it's not possible to compute the correct 3D translation, but we are able to estimate the parameters (U', V', W), which shows the same angular characteristics in 2D as the true 3D translation (U, V, W). Bruss and Horn give a closed form solution to this problem for the translation-only case.

Recovering camera rotation. Bruss and Horn also outline how to solve for rotation, but give limited details. We implement our own estimation of rotations (A, B, C) and translation as a nested optimization:

$$\hat{M} = \underset{A,B,C,U',V',W}{\arg\min} \left[\min_{U,V,W} \sum_i \|\boldsymbol{e}_i\left(\boldsymbol{v}_i, A, B, C, U', V', W\right)\| \right]. \tag{14}$$

Given (A, B, C) one can compute the flow vectors $\boldsymbol{m_r}$ describing the rotational motion field of the observed flow, one can subtract off the rotation since it does not depend on scene geometry: $\hat{\boldsymbol{v}}_{\boldsymbol{t}} = \boldsymbol{v} - \hat{\boldsymbol{m}}_{\boldsymbol{r}}(\hat{A}, \hat{B}, \hat{C})$. Subtracting the rotation (A, B, C) from the observed flow reduces the optimization to the translation only case. We solve the optimization over the rotation parameters A, B, C by using Matlab's standard gradient descent optimization, while calling the Bruss and Horn closed form solution for the translation variables given the rotational variables as part of the internal function evaluation. Local minima are a concern, but since we are estimating camera motion between two video frames, the rotation is almost always small and close to the optimization's starting point.

3.3 Initialization: Segmenting the first frame

The goals of the initialization are a) estimating translation parameters (U', V', W) and the rotation (A, B, C) of the motion of static environment due to camera motion, b) the estimated set of parameters (U', V', W) form an angle field M corresponging to the observed flow c) finding pixels whose flow is consistent with M, and d) assigning inconsistent groups of contiguous pixels to additional angle fields. Bruss and Horn's method was not developed to handle scenes with multiple different motions, and so large or fast-moving objects can result in poor motion estimates (Figure 8).

Constrained RANSAC. To address this problem we use a modified version of RANSAC [33] to robustly estimate motion of static environment (Figure 6). We use 10 random SLIC superpixels [34][1] to estimate camera motion (Section 3.2). We modify the standard RANSAC procedure to force the algorithm to choose three of the 10 patches from the image corners, because image corners are prone to errors due to a misestimated camera rotation. Since the

[1] We use the http://www.vlfeat.org/api/slic.html code with regionSize=20 and regularizer=0.5.

Bruss and Horn error function (Equation 14) does not penalize motions in a direction opposite of the predicted motion, we modify it to penalize these motions appropriately (details in Supp. Mat.). 5000 RANSAC trials are run, and the camera motion resulting in the fewest outlier pixels according to the *modified Bruss-Horn* (MBH) error is retained, using a threshold of 0.1.

Input: video with n frames
Output: binary motion segmentation

for $t \leftarrow$ *1* ***to*** $n-1$ **do**
  compute optical flow from frame t to frame $t+1$
  if *first frame* **then**
    foreach *RANSAC iteration* **do**
      find best set of translation parameters (U', V', W) for 10 random patched (3 in corners)
      retain best angle field for the static environment M_k
    end
    $p(M) \leftarrow$ segment MBH error image into k comp. using Otsu's method
  else
    $p(M) \leftarrow$ propagate posterior $p(M \mid \boldsymbol{v_t})$
    find (U', V', W) and rotation (A, B, C) of static environment using gradient descent
    foreach *flow vector* $\boldsymbol{v}$ **do**
      $\boldsymbol{v_t} = \boldsymbol{v} - \boldsymbol{m_r}(A, B, C)$
    end
  end
  for $j \leftarrow 1$ ***to*** k **do**
    compute angle field M_j of motion component j
    foreach *flow vector* $\boldsymbol{v_t}$ **do**
      $p(\theta_{\boldsymbol{v_t}} \mid M_j, \|\boldsymbol{v_t}\|) \leftarrow \mathcal{V}(\theta_{\boldsymbol{v_t}}; \mu = \theta^j_{\boldsymbol{m_t}}, \kappa = a\|\boldsymbol{v_t}\|^b)$
    end
  end
  foreach *flow vector* $\boldsymbol{v_t}$ **do**
    $p(M_{k+1}) \leftarrow \frac{1}{k+1}$
    $p(\theta_{\boldsymbol{v_t}} \mid M_{k+1}, \|\boldsymbol{v_t}\|) \leftarrow \frac{1}{2\pi}$
    normalize $p(M_j)$ that they sum up to $1 - p(M_{k+1})$
    $p(M \mid \boldsymbol{v_t}) \leftarrow p(\theta_{\boldsymbol{v_t}} \mid M, \|\boldsymbol{v_t}\|) \cdot p(M)$
  end
  given the posteriors $p(M \mid \boldsymbol{v_t})$ assign every pixel one of two labels: static environment or moving objects
end

Fig. 5: A causal motion segmentation algorithm

Otsu's Method. While using the RANSAC threshold on the MBH image produces a good set of pixels to estimate the motion of the static environment due to camera motion, the method often excludes some pixels that should be included

in the motion component of static environment. We use Otsu's method [35] to separate the MBH image into a region of low error (static environment) and high error: (1) Use Otsu's threshold to divide the errors, minimizing the intraclass variance. Use this threshold to do a binary segmentation of the image. (2) Find the connected component C with highest average error. Remove these pixels ($I \leftarrow I \setminus C$), and assign them to an additional angle field M. These steps are repeated until Otsu's *effectiveness* parameter is below 0.6.

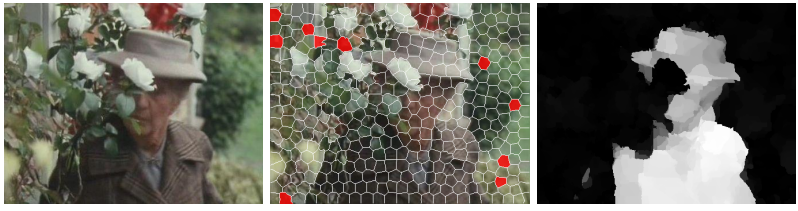

Fig. 6: **RANSAC procedure.** The result of our RANSAC procedure is to find image patches of the static environment. Notice that none of the patches are on the person moving in the foreground. Also notice that we force the algorithm to pick patches in three of the four image corners (a "corner" is 4% of the image). The right figure shows the negative log likelihood of the static environment.

4 Experiments

Several motion segmentation benchmarks exist, but often a clear definition of what people intend to segment in ground truth is missing. The resulting inconsistent segmentations complicate the comparison of methods. We define motion segmentation as follows.

(I) Every pixel is given one of **two labels**: static environment or moving objects.

(II) If only part of an object is moving (like a moving person with a stationary foot), the **entire object** should be segmented.

(III) **All freely moving objects** (not just one) should be segmented, but nothing else. We do not considered tethered objects such as trees to be freely moving.

(IV) Stationary objects are not segmented, even when they moved before or will move in the future. We consider segmentation of previously moving objects to be *tracking*. Our focus is on segmentation by motion analysis.

Experiments were run on two previous data sets and our new camouflaged animals videos. The first was the Berkeley Motion Segmentation (BMS-26) database [2, 3] (Figure 9, rows 5,6). Some BMS videos have an inconsistent definition of ground truth from both our definition and from the other videos in the benchmark. An example is *Marple10* whose ground truth segments a wall in the foreground as a moving object (see Figure 7). While it is interesting to

use camera motion to segment static objects (as in [36]), we are addressing the segmentation of objects that are moving differently than the static environment, and so we excluded ten such videos from our experiments (see Supp. Mat.). The second database used is the Complex Background Data Set [4], which includes significant depth variation and also significant amounts of camera rotation (Figure 9, rows 3,4). We also introduce the Camouflaged Animals Data Set (Figure 9, rows 1,2) which will be released at camera-ready time. These videos were ground-truthed every 5th frame. See Supp. Mat. for more.

Fig. 7: **Bad ground truth.** Some BMS-26 videos contain significant ground truth errors, such as this segmentation of the foreground wall, which is clearly not a moving object.

Setting von Mises parameters. There are two parameters a and b that affect the von Mises concentration $\kappa = a\|\boldsymbol{m_t}\|^b$. To set these parameters for each video, we train on the remaining videos in a leave-one-out paradigm, maximizing over the values $0.5, 1.0, 2.0, 4.0$ for multiplier parameter a and the values $0, 0.5, 1, 2$ for the exponent parameter b. Cross validation resulted in the selection of the parameter pair $(a = 4.0, b = 1.0)$ for most videos, and we adopted these as our final values.

		Keuper [18]	**Papaz. [17]**	**Frag. [20]**	**Zama. [16]**	**Naray. [4]**	**ours**
Camouflage	MCC	**0.4305**	0.3517	0.1633	0.3354	-	**0.5344**
	F	**0.4379**	0.3297	0.1602	0.3007	-	**0.5276**
BMS-26	MCC	0.6851	0.6112	**0.7187**	0.6349	-	**0.7576**
	F	**0.7306**	0.6412	0.7276	0.6595	0.6246	**0.7823**
Complex	MCC	0.4752	**0.6359**	0.3257	0.3661	-	**0.7491**
	F	0.4559	**0.6220**	0.3300	0.3297	0.3751	**0.7408**
Total avg.	MCC	**0.5737**	0.5375	0.4866	0.5003	-	**0.6918**
	F	**0.5970**	0.5446	0.4911	0.4969	-	**0.6990**

Table 1: **Comparison to state-of-the-art.** Matthew's correlation coefficient and F-measure for each method and data set. The "Total avg." numbers average across all valid videos.

Results. In Tab. 1, we compare our model to five different state-of-the-art methods [4, 16–18, 20]. We compared against methods for which either code was available or that had results on either of the two public databases that we used. However, we excluded some methods (such as [19]), as their published results were less accurate than [18], to whom we compared.

Some authors have scored algorithms using the number of correctly labeled pixels. However, when the moving object is small, a method can achieve a very high score simply by segmenting the entire video with the label static environment. The F-measure is also not symmetric with respect to a binary segmentation, and is not well-defined when a frame contains no moving pixels. Matthew's Correlation Co-efficient (MCC) handles both of these issues, and is recommended for scoring such binary classification problems when there is a large imbalance between the number of pixels in each category [37]. However, in order to enable comparison with [4], and to allow easier comparison to other methods, we also included F-measures. Table 1 shows the highest average accuracy per data set in **green** and the second best in **blue**, for both the F-measure and MCC. We were not able to obtain code for Narayana et al. [4], but reproduced F-measures directly from their paper. The method of [20] failed on several videos (only in the BMS data set), possibly due to the length of these videos. In these cases, we assigned scores for those videos by assigning all pixels to static environment.

Our method outperforms all other methods by a large margin, on all three data sets, using both measures of comparison.

5 Analysis and Conclusions

Conditioning our angle likelihood on the flow magnitude is an important factor in our method. Table 2 shows the detrimental effect of using a constant von Mises concentration κ instead of one that depends upon $\|\boldsymbol{m_t}\|$. In this experiment, we set the parameter b which governs the dependence of κ on $\|\boldsymbol{m_t}\|$ to 0, and set the value of κ to maximize performance. Even with the optimum constant κ, the drop in performance was 7%, 5%, and 22% across three data sets.

We also show the consistent gains stemming from our constrained RANSAC initialization procedure. In this experiment, we segmented the first frame of video without rejecting any pixels as outliers. In some videos, this had little effect, but sometimes the effect was large, as shown in Figure 8 – here the estimated M is the best fit for the car instead of static environment.

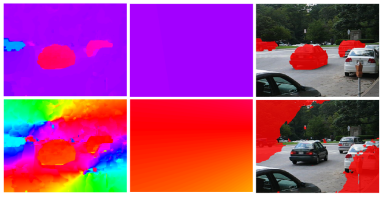

Fig. 8: **RANSAC vs no RANSAC.** Top row: robust initialization with RANSAC. Bottom row: using Bruss and Horn's method directly on the entire image. Left to right: flow angles of observed translational flow, angle field M of static environment and segmentation.

	final	**constant κ**	**no RANSAC**
BMS-26	**0.7576**	0.6843	0.6450
complex	**0.7491**	0.7000	0.5757
camouflage	**0.5344**	0.3128	0.5176

Table 2: **Effect of RANSAC and variable κ.**

The method by Keuper et al. [18] performs fairly well, but often makes errors in segmenting rigid parts of the foreground near the observer. This can be seen in the third and fourth rows of Figure 9, which shows sample results from the Complex Background Data Set. In particular, note that Keuper et al.'s method segments the tree in the near foreground in the third row and the wall in the near foreground in the fourth row. The method of Fragkiadaki et al., also based on trajectories, has similar behavior. These methods in general seem to have difficulty with high variability in depth.

Another reason for our good results may be that we are directly using the perspective projection equations to analyze motion, as has been advocated by Horn [27], rather than approximations based on projective geometry. Code is available: `http://vis-www.cs.umass.edu/motionSegmentation/`.

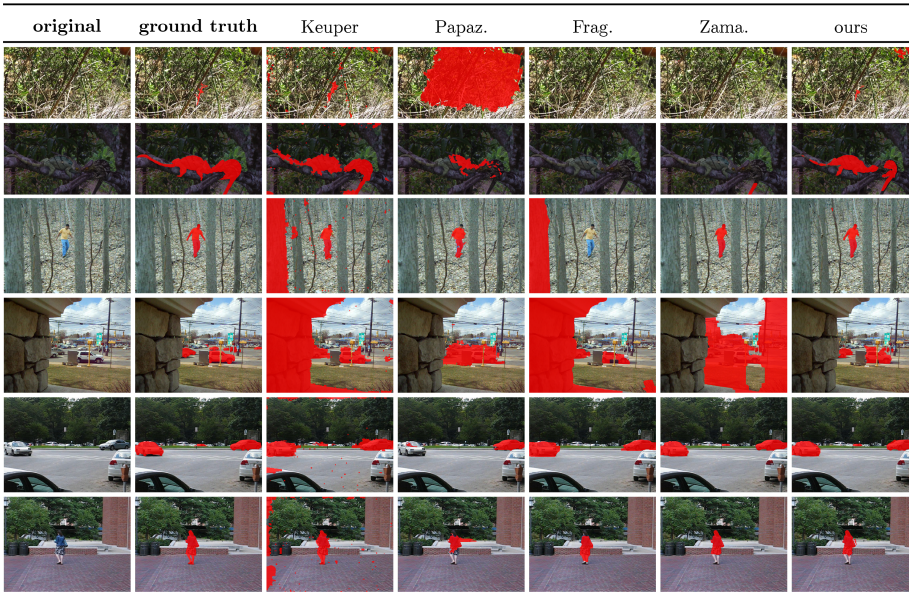

Fig. 9: **Sample results** Left to right: original image, ground truth, [18], [17], [20] [16] and our binary segmentations. Rows 1-2: sample results on the Animal Camouflage Data Set (chameleon and stickinsect). Rows 3-4: sample results on Complex Background (traffic and forest). Rows 5-6: sample results on BMS-26 (cars5 and people1).

References

1. Torr, P.H.: Geometric motion segmentation and model selection. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **356**(1740) (1998) 1321–1340
2. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: ECCV. (2010)
3. Tron, R., Vidal, R.: A benchmark for the comparison of 3-d motion segmentation algorithms. In: CVPR. (2007)
4. Narayana, M., Hanson, A., Learned-Miller, E.: Coherent motion segmentation in moving camera videos using optical flow orientations. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE (2013) 1577–1584
5. Grundmann, M., Kwatra, V., Han, M., Essa, I.: Efficient hierarchical graph-based video segmentation. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 2141–2148
6. Lezama, J., Alahari, K., Sivic, J., Laptev, I.: Track to the future: Spatio-temporal video segmentation with long-range motion cues. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2011)
7. Kumar, M.P., Torr, P.H., Zisserman, A.: Learning layered motion segmentations of video. International Journal of Computer Vision **76**(3) (2008) 301–319
8. Irani, M., Rousso, B., Peleg, S.: Computing occluding and transparent motions. International Journal of Computer Vision **12** (1994) 5–16
9. Ren, Y., Chua, C.S., Ho, Y.K.: Statistical background modeling for non-stationary camera. Pattern Recognition Letters **24** (2003) 183 – 196
10. Sheikh, Y., Javed, O., Kanade, T.: Background subtraction for freely moving cameras. In: IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009, IEEE (2009) 1219–1225
11. Elqursh, A., Elgammal, A.M.: Online moving camera background subtraction. In: ECCV. (2012)
12. Ochs, P., Brox, T.: Higher order motion models and spectral clustering. In: CVPR. (2012)
13. Kwak, S., Lim, T., Nam, W., Han, B., Han, J.H.: Generalized background subtraction based on hybrid inference by belief propagation and Bayesian filtering. In: ICCV. (2011)
14. Rahmati, H., Dragon, R., Aamo, O.M., Van Gool, L., Adde, L.: Motion segmentation with weak labeling priors. In: Pattern Recognition. Springer (2014) 159–171
15. Jain, S.D., Grauman, K.: Supervoxel-consistent foreground propagation in video. In: Computer Vision–ECCV 2014. Springer (2014) 656–671
16. Zamalieva, D., Yilmaz, A., Davis, J.W.: A multi-transformational model for background subtraction with moving cameras. In: Computer Vision–ECCV 2014. Springer (2014) 803–817
17. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE (2013) 1777–1784
18. Keuper, M., Andres, B., Brox, T.: Motion trajectory segmentation via minimum cost multicuts. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 3271–3279
19. Taylor, B., Karasev, V., Soatto, S.: Causal video object segmentation from persistence of occlusions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 4268–4276

20. Fragkiadaki, K., Zhang, G., Shi, J.: Video segmentation by tracing discontinuities in a trajectory embedding. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 1846–1853
21. Sawhney, H.S., Guo, Y., Asmuth, J., Kumar, R.: Independent motion detection in 3d scenes. In: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on. Volume 1., IEEE (1999) 612–619
22. Dey, S., Reilly, V., Saleemi, I., Shah, M.: Detection of independently moving objects in non-planar scenes via multi-frame monocular epipolar constraint. In: Computer Vision–ECCV 2012. Springer (2012) 860–873
23. Namdev, R.K., Kundu, A., Krishna, K.M., Jawahar, C.V.: Motion segmentation of multiple objects from a freely moving monocular camera. In: International Conference on Robotics and Automation. (2012)
24. Csurka, G., Bouthemy, P.: Direct identification of moving objects and background from 2d motion models. In: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on. Volume 1., IEEE (1999) 566–571
25. Sharma, R., Aloimonos, Y.: Early detection of independent motion from active control of normal image flow patterns. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **26**(1) (1996) 42–52
26. Elhamifar, E., Vidal, R.: Sparse subspace clustering. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 2790–2797
27. Horn, B.K.: Projective geometry considered harmful (1999)
28. Ogale, A.S., Fermüller, C., Aloimonos, Y.: Motion segmentation using occlusions. Pattern Analysis and Machine Intelligence, IEEE Transactions on **27**(6) (2005) 988–992
29. Horn, B.: Robot vision. MIT Press (1986)
30. Bruss, A.R., Horn, B.K.: Passive navigation. Computer Vision, Graphics, and Image Processing **21**(1) (1983) 3–20
31. Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 2432–2439
32. Narayana, M., Hanson, A., Learned-Miller, E.G.: Background subtraction: separating the modeling and the inference. Machine vision and applications **25**(5) (2014) 1163–1174
33. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6) (1981) 381–395
34. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. Pattern Analysis and Machine Intelligence, IEEE Transactions on **34**(11) (2012) 2274–2282
35. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics **9** (1979) 62–66
36. Wang, J.Y., Adelson, E.H.: Representing moving images with layers. Image Processing, IEEE Transactions on **3**(5) (1994) 625–638
37. Powers, D.M.: Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. Technical Report Technical Report SIE-07-001, Flinders University, Adelaide (2007)