Inlier Estimation for Moving Camera Motion Segmentation

Xuefeng Liang, Cuicui Zhang, Takashi Matsuyama

IST, Graduate School of Informatics, Kyoto University, Japan.

Abstract. In moving camera videos, motion segmentation is often performed on the optical flow. However, there exist two challenges: 1) Camera motions lead to three primary flows in optical flow: translation, rotation, and radial flow. They are not all solved in existing frameworks under Cartesian coordinate system; 2) A moving camera introduces 3D motion, the depth discontinuities cause the motion discontinuities that severely confuse the motion segmentation. Meanwhile, the mixture of the camera motion and moving objects' motions make indistinctness between foreground and background. In this work, our solution is to find a low order polynomial to model the background flow field due to its coherence. To this end, we first amend the Helmholts-Hodge Decomposition by adding coherence constraints, which can handle translation, rotation, and radial flow fields under a unified framework. Secondly, we introduce an Incoherence Map and a progressive Quad-Tree partition to reject moving objects and motion discontinuities. Finally, the low order polynomial is achieved from the rest flow samples on two potentials in HHD. We present results on more than twenty videos from four benchmarks. Extensive experiments demonstrate a better performance in dealing with challenging scenes with complex backgrounds. Our method improves the segmentation accuracy of state-of-the-art by $10\% \sim 30\%$.

1 Introduction

With rapid increase of mobile cameras (handhold camera, wearable camera, etc), video analysis faces more challenges. One of them is that appearance motions are no longer simple in such scenes where multiple objects could move independently, in addition to the camera motion. It is named 3D motion. A common scheme in 3D motion segmentation is to use optical flow or trajectories as a cue. As optical flow can be directly used for clustering or to compensate for the camera motion, the pixelwise model are often used for segmentation [1] [2] [3] [4].

However, here are two major drawbacks of using optical flow. 1) Camera motions in 3D scene cause three primary motion flows in optical flow: translation, rotation, and radial flow. The well accepted interpretation of the flow vector is based on Cartesian coordinate system with two bases x, y. Then, a motion vector in 2D is denoted by u, v. For translation, it can be interpreted as an invariant u/v that depends on depth and camera motion only. But for rotation and radial flow, the u/v changes with x, y changing on image plan too. Then, there is no

invariant to interpret them under the Cartesian coordinate system. For example, Fig.1.(b) and Fig.4.(b) involve the radial flow and rotation, respectively. The direction and magnitude of flow vectors are changing w.r.t the change of x, y. To our best knowledge, no scheme could well handle all these three flows based on dense flow field. 2) Optical flow depends on the object's distance from the camera. A varied depth causes a different magnitude of the flow. This may make a clustering algorithm to label backgrounds at different depths as separate objects although they are static in the real world. In Fig.1.(b), there is a significant motion discontinuity between the stopped car and the far away background. To handle motion discontinuities, other works need auxiliary information (e.g. color, edge energy) to merge small segments into one as a post-processing.

As the flow filed of the static background (inlier) in optical flow is caused by the camera motion, it should be globally coherent. But, existing optical flow algorithms give much error at the depth discontinuity, and make those flows incoherent with the camera motion. For a moving object in 3D scene, its flow field (outlier), however, is caused by not only its own motion but also the camera motion. So, an outlier consists of both incoherent and coherent flows w.r.t. the camera motion. Thus, the coherence in optical flow can be a cue for 3D motion segmentation. In this work, our object is to find an appropriate polynomial for inlier modeling according to its coherence, which requires no prior knowledge and post-processing. To this end, the first challenge is how to put three primary motion flows into one scheme. Helmholtz-Hodge decomposition (HHD) was initially developed to characterize the rotation and radial flow by curl-divergence regularization. HHD can decompose an arbitrary motion field into *curl-free* and *divergence-free* components through finding their unique corresponding scalar and vector potentials no matter the motion is coherent or not. To ensure the coherence, we amend conventional HHD by adding two constraints: piece-wise smoothness, and global minimization. Nevertheless, the obtained two potentials consist of the major coherent inlier and a little coherent outliers. To better estimate inlier, we introduce an Incoherence Map (IM) by subtracting the projection of two potentials from optical flow. It intuitively depicts the outliers and motion discontinuities. Moreover, a progressive Quad-Tree partition is proposed for precisely labeling outliers and motion discontinuities on IM, and rejecting them from two potentials. Therefore, outliers is completely excluded in inlier estimation. The motion discontinuities are also excluded, but do not affect the coherent flows at the fields with different depths. Afterwards, our object can be achieved by approximating the low order polynomials using rest samples on two potentials.

Other approaches for 3D motion segmentation can be categorized as (1) using optical flow, and (2) trajectories clustering. Chen and Bajic [1] proposed an outlier rejection filter that explicitly filters motion vectors by checking their similarity in a pre-defined window. Chen and Bajic [2], and Qian and Bajic [3] proposed a joint global motion estimation, which iteratively update the inlier model by segmenting outliers out. Although these methods have achieved great progress in dealing with independent motions, they are very likely to over-segment ob-

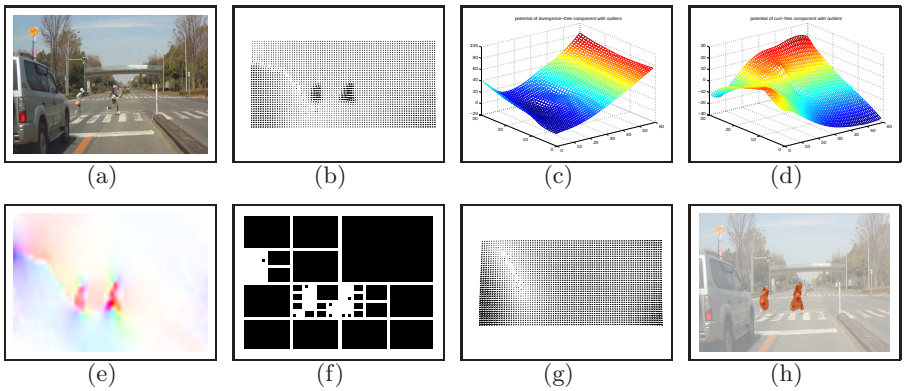


Fig. 1. (a) A traffic scene where two kids are running from left to right. A car stops at the left side. The camera is moving towards the stop line. (b) The original optical flow. Besides two outliers, here exists a significant flow discontinuity between the left side car and far away background. (c) The potential of divergence-free component. (d) The potential of curl-free component. (e) Incoherence Map *IM*. (f) Progressive Quad-Tree Partition on IM to reject the outliers, motion discontinuities, and noise. (g) The estimated inlier. It has been rather coherent. (h) Result of 3D motion segmentation. It detects outliers only.

jects due to the motion bias introduced by camera motion [2]. Narayana et al. [4] proposed a method using the direction of motion flow only. It works well for 2D translation, but has difficulty with rotation and radial flow. Brox and Malik [5] segment trajectories by computing the pairwise distances between all trajectories and finding a low-dimensional embedding using spectral clustering. Later, Ochs and Brox [6] improved the spectral clustering by using higher order interactions that consider triplets of trajectories. Elqursh and Elgammal [7] proposed an online extension of spectral clustering by considering trajectories from multiple frames. But, they require a post-processing for merging. Kwak et al. [8] use a Bayesian filtering framework that combines block-based color appearance models with separate motion models for segmentation. However, they require a special initialization procedure in the first frame.

In contrast, our method is a frame-to-frame scheme, requires neither trajectories from multiple frames nor special initialization and post-processing. In additional, it handles all three primary flows in one scheme, and works on a rather wide bank of videos.

2 Modeling inlier of optical flow

2.1 Models of optical flow and 3D motion segmentation

Let X, Y, Z denote the horizontal, vertical and depth axes in Cartesian coordinate of a real world, and let x, y denote the corresponding coordinates in the image plane. The image plane is located at the focal length: $Z' = f$. In 3D scene, the camera motion has two components: a translation $T = (T_X, T_Y, T_Z)$ and a

rotation $R = (R_X, R_Y, R_Z)$. They are always coherent (continues and smooth in both direction and magnitude) in a short time interval $\triangle t$. Then, the resulting 2D optical flow u and v in the x and y image axes are [9]

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} -f(\frac{T_X}{Z} + R_Y) + x\frac{T_Z}{Z} + yR_Z - x^2\frac{R_Y}{f} + xy\frac{R_X}{f} \\ -f(\frac{T_Y}{Z} - R_X) + y\frac{T_Z}{Z} - xR_Z + y^2\frac{R_X}{f} - xy\frac{R_Y}{f} \end{bmatrix}. \tag{1}$$

Beside constant f, camera motion T and R, and image coordinates x and y, it is noted from Eq.(1) that u and v are functions with depth Z too. Thus, an ideal optical flow field ($\mathcal{OF}$) in 3D scene is a collection of 2D $\mathcal{OF}$s of n static planar surfaces in background (named *inlier*, $\mathcal{F}_{in}$) and $\mathcal{OF}$s of m foreground moving objects (named *outlier*, $\mathcal{F}_{out}$). We can formulate them as

$$\mathcal{OF} = \mathcal{F}_{in} + \mathcal{F}_{out}, \tag{2}$$

where $\mathcal{F}_{in} = [\mathcal{F}_{in}(B_1)\ \mathcal{F}_{in}(B_2) \dots \mathcal{F}_{in}(B_n)]$, $\mathcal{F}_{out} = [\mathcal{F}_{out}(O_1)\ \mathcal{F}_{out}(O_2) \dots \mathcal{F}_{out}(O_m)]$, B_i denotes a static planar surface and O_j denotes a moving object. Please note that $\mathcal{F}_{out}(O_j)$ is incoherent with $\mathcal{F}_{in}$ because its own motion does not match the camera motion.

Theoretically speaking, each $\mathcal{F}_{in}(B_i)$ or $\mathcal{F}_{out}(O_j)$ could be approximated by a polynomial. Thus, entire $\mathcal{OF}$ could be also approximated by a high order polynomial $\mathcal{P}$.

$$\mathcal{OF} = [\mathcal{F}_{in}(B_1) \dots \mathcal{F}_{in}(B_n)\ \mathcal{F}_{out}(O_1) \dots \mathcal{F}_{out}(O_m)] \approx \mathcal{P}, \tag{3}$$

where high order is required due to the outliers, motion discontinuities caused by depth discontinuities, and noise.

To segment 3D motion, many studies tried to model outliers directly. Moving objects, however, can be either rigid or non-rigid, outlier modeling is a nontrivial task. Instead of complex modeling and auxiliary constrains, we model inlier by a general approach in this work other than modeling outliers. As the inlier is caused by camera motion, each $\mathcal{F}_{in}(B_i)$ must be coherent, and can be modeled by a simple polynomial $\mathcal{P}_i$. But Eq.(1) shows $\mathcal{F}_{in}(B_i)$ is a function with Z in translation and radial flow fields. Modeling $\mathcal{F}_{in}$ is still difficult.

Most optical flow algorithms share a common assumption of local motion smoothness, and apply an optimization to minimize the global error. The difference among them only focuses on implementations of the optimization. This strategy is ideally designed for the motion of one planar surface. But, it is also applied on the object's boundaries, where depths vary, because algorithms do not know where the depth discontinuities are. For inlier $\mathcal{F}_{in}$, at the place where depth variation is not significant, optical flow algorithms give rather smooth flow field and connect motion fields of two adjacent planar surfaces. By the same token, algorithms often give much errors at the place having significant depth discontinuity. These incorrect flows are incoherent with camera motion, but are minority in $\mathcal{F}_{in}$. Then, $\mathcal{F}_{in}$ can be reformulated as

$$\mathcal{F}_{in} = a_1\mathcal{F}_{in}^{coherent} + b_1\mathcal{F}_{in}^{incoherent}, \qquad a_1 >> b_1, \tag{4}$$

where a and b are quantity coefficients.

The primary reason of the moving objects standing out in a moving camera video is that their motions are incoherent with camera motion and result in significant variances on optical flow field $\mathcal{OF}$. Besides, $\mathcal{F}_{out}$ is partially caused by camera motion as well, which is coherent with $\mathcal{F}_{in}$, but is minority. Now, $\mathcal{F}_{out}$ can be reformulated as

$$\mathcal{F}_{out} = a_2\mathcal{F}_{out}^{coherent} + b_2\mathcal{F}_{out}^{incoherent}, \qquad b_2 >> a_2. \tag{5}$$

One major difficulty in 3D motion segmentation is induced by the mixed flow field of camera motion and moving objects' motions. These dependent motions lead to indistinctness of the difference between inlier $\mathcal{F}_{in}$ and outliers $\mathcal{F}_{out}$. To segment outliers from inlier, our object, therefore, becomes finding an appropriate polynomial

$$\mathcal{P}' \approx a_1\mathcal{F}_{in}^{coherent} + a_2\mathcal{F}_{out}^{coherent}, \tag{6}$$

which rejects incoherence ($b_1\mathcal{F}_{in}^{incoherent} + b_2\mathcal{F}_{out}^{incoherent}$) in inlier and outliers.

2.2 Three primary motion flows and their potentials

Equation (1) shows motion vectors in $\mathcal{OF}$ caused by camera motion can be decomposed into two components: V_T and V_R representing camera translation and rotation, respectively.

$$V_T = \left[\frac{xT_Z - fT_X}{Z}, \frac{yT_Z - fT_Y}{Z}\right]^T,\ V_R = \begin{bmatrix} yR_Z - x^2\frac{R_Y}{f} + xy\frac{R_X}{f} - fR_Y \\ -xR_Z + y^2\frac{R_X}{f} - xy\frac{R_Y}{f} + fR_X \end{bmatrix}.$$

To simplify analysis, motion vectors caused by camera motion can be further decomposed into three primary components:

1. $V_{T_{\mathbb{X}}}$ caused by camera translation on image plan $\mathbb{X}(x, y)$:

$$V_{T_{\mathbb{X}}} = \left[-\frac{fT_X}{Z}, -\frac{fT_Y}{Z}\right],\ \arctan(V_{T_{\mathbb{X}}}) = \frac{T_X}{T_Y},\ |V_{T_{\mathbb{X}}}| = \frac{f}{Z}\sqrt{T_X^2 + T_Y^2}. \tag{7}$$

   It is noted that the direction of flow is independent on the depth Z, and the magnitude is inversely proportional to the depth Z. Eq.(7) indicates that if the depth variation is not significant in 3D scene (e.g. the static background is rather far from the camera), the coherence of both direction and magnitude is preserved. If background is pretty close to the camera, the depth discontinuity will lead to motion discontinuity that is incoherent with camera motion.
2. V_{T_Z} caused by camera translation along Z axis. It presents a radial flow field with the origin at the focus-of-expansion.

$$V_{T_Z} = \left[\frac{xT_Z}{Z}, \frac{yT_Z}{Z}\right], \quad \arctan(V_{T_Z}) = \frac{x}{y}, \quad |V_{T_Z}| = \frac{T_Z}{Z}\sqrt{x^2 + y^2}. \tag{8}$$

It is noted that the flow direction is dependent on neither the depth Z nor camera motion, but only determined by image plan coordinates x and y. The magnitude is also inversely proportional to the depth Z. Analogically, the coherent flow field is maintained at the place having less depth variation. But, the incoherence occurs at the place having significant depth discontinuity. Please see Fig.1.(b) as an example.

3. V_R caused by camera rotation along an axis parallel with the camera optical axis Z. Please note that camera rotation, which rotates along image axes x or y in a short time interval $\triangle t$, could be simulated as a translation. In this work, we consider camera rotation perpendicular to the image plan only.

$$V_R = [yR_Z, -xR_Z], \quad \arctan(V_R) = -\frac{y}{x}, \quad |V_R| = R_Z\sqrt{x^2+y^2}. \tag{9}$$

It is noted that both direction and magnitude of V_R are independent on depth Z, but dependent on image plan coordinates x and y. This indicates that a 3D motion segmentation can reduce to a 2D motion segmentation while camera solely rotates perpendicular to the image plan. Obviously, the coherence of flow field is preserved.

Consequently, an arbitrary optical flow field $\mathcal{OF}$ can be represented by a combination of above three primary flows as:

$$\mathcal{OF} = \alpha V_{T_{\mathbb{X}}} + \beta V_{T_Z} + \gamma V_R,$$

where α, β and γ are quantity coefficients.

Analogically, Prof. Helmholtz explained that the motion of a volume element in 3D space consists of: *1) expansion or contraction, 2) rotation, and 3) translation. The expansion/contraction (radial flow field) can be represented as the gradient of a scalar potential function because it is irrotational. The rotation can be represented as the curl of a vector potential function since it is incompressible. Translation, however, being neither compressible nor rotational can be represented as either the gradient of a scalar potential, or the curl of a vector potential*[10]. It is named by *Helmholtz-Hodge Decomposition* (HHD) [11]. According to HHD, any flow field in our work consists of two components:

1. **Curl-free component** representing divergence (radial flow) and translation because they are irrotational.

$$\theta = \nabla \cdot \mathcal{OF} = V_{T_Z} + V_{T_{\mathbb{X}}}$$

2. **Divergence-free component** representing curl (rotation) and translation because they are incompressible.

$$\overrightarrow{\omega} = \nabla \times \mathcal{OF} = V_{T_R} + V_{T_{\mathbb{X}}}$$

Go a step further, curl-free and divergence-free components can be expressed as the curl of a vector potential $\overrightarrow{W}$ and the gradient of a scalar potential E,

$$\mathcal{OF} = \nabla E + \nabla \times \overrightarrow{W}. \tag{10}$$

where $E(x) = -\frac{1}{4\pi}\int \frac{\theta(x')}{|x-x'|}dx'$, $\overrightarrow{W}(x) = -\frac{1}{4\pi}\int \frac{\overrightarrow{\omega}(x')}{|x-x'|}dx'$, and $x \in \mathbb{R}^3$. However, θ and $\overrightarrow{\omega}$ are what we expect. In reality, E and $\overrightarrow{W}$ are computed by energy minimization,

$$\min F(E) = \min \int (\mathcal{OF} - \nabla E)^2, \ \min G(\overrightarrow{W}) = \min \int (\mathcal{OF} - \nabla \times \overrightarrow{W})^2.$$

Theoretically speaking, HHD can decompose an arbitrary motion flow field into curl-free component ∇E and divergence-free component $\nabla \times \overrightarrow{W}$, no matter it is coherent or not. But, we expect coherent potentials for approximating the coherent flow field caused by camera motion, see Eq.(6). To ensure the coherence, our work amended the conventional HHD by adding the first constraint: 1) piece-wise smooth $\mathcal{S}$. Meanwhile, we make an assumption $\mathcal{F}_{in} > \mathcal{F}_{out}$, and add the second constraint: 2) global minimization at entire motion flow field Ω, which ensures the optimization to minimize the inlier, other than the outliers.

$$\begin{aligned} \arg\min_{\mathcal{S}} F(E) &= \arg\min_{\mathcal{S}} \int_{\Omega} (\mathcal{OF} - \nabla E)^2 d\Omega, \\ \arg\min_{\mathcal{S}} G(\overrightarrow{W}) &= \arg\min_{\mathcal{S}} \int_{\Omega} (\mathcal{OF} - \nabla \times \overrightarrow{W})^2 d\Omega. \end{aligned} \tag{11}$$

Please see Fig.1(c) and (d), they are the divergence-free and curl-free potentials respectively. Therefore, these two potential E and $\overrightarrow{W}$ could be our object: coherent surfaces which can be formulated by low order polynomial $\mathcal{P}'$.

2.3 Incoherence map and incoherence labeling

Due to the global optimization and piece-wise smooth constraint, our amended implementation keeps most coherent flow ($95\% \sim 99\%$) into two potentials, but definitely rejects incoherent flow $\mathcal{F}_{in}^{incoherent}$ and $\mathcal{F}_{out}^{incoherent}$. Then, the majority of outliers, motion discontinuities and noise, which are not decomposed into ∇E and $\nabla \times \overrightarrow{W}$, will rest in a remainder. Please note that two potentials still contain a small quantity of $\mathcal{OF}_{out}^{coherent}$. They cannot be directly used for the inlier estimation. Thus, we use this remainder to draw an *Incoherence Map* (IM) to label incoherent flows in $\mathcal{OF}$.

First, we estimate the coherent flow field presented in the curl-free and divergence-free components by a linear combination as:

$$\mathcal{V} = \alpha(\nabla E) + \beta(\nabla \times \overrightarrow{W}), \tag{12}$$

where $\mathcal{V} \subseteq \mathcal{OF}$, α and β are quantity coefficients that indicate how much ∇E and $\nabla \times \overrightarrow{W}$ are involved in $\mathcal{OF}$. We use a distance to determine α and β.

$$d_\theta = \int \frac{|\mathcal{OF} - \nabla E|}{|\mathcal{OF}|}, \qquad d_\omega = \int \frac{|\mathcal{OF} - \nabla \times \overrightarrow{W}|}{|\mathcal{OF}|}, \tag{13}$$

where d_θ represents the distance between the curl-free component and the optical flow field, and d_ω represents the distance between the divergence-free component and the optical flow field. Then, α and β are determined as following:

- While $d_\theta < 0.5$ and $d_\omega > 0.5$, it implies the optical flow looks more like a radial flow field (curl-free component).
- While $d_\theta > 0.5$ and $d_\omega < 0.5$, it implies the optical flow looks more like a rotation field (divergence-free component).
  Under above two cases,

$$\alpha = \frac{d_\theta}{d_\theta + d_\omega}, \qquad \beta = \frac{d_\omega}{d_\theta + d_\omega}.$$

- While $d_\theta < 0.5$ and $d_\omega < 0.5$, it means both components are similar to the optical flow, and implies a translation. Then,

$$\alpha = 0.5, \qquad \beta = 0.5.$$

Afterwards, we can draw Incoherence Map from remainder by

$$IM = \mathcal{OF} - \alpha(\nabla E) - \beta(\nabla \times \overrightarrow{W}). \tag{14}$$

Please see Fig.1.(e), it clearly reveals the outliers, depth discontinuities and computation error.

IM makes outliers and motion discontinuities labeling much easier. However, Eq.(12) and Eq.(14) show that IM consists of a small portion of inlier as well. To ensure the accurate labeling, we introduce a *Progressive Quad-Tree Partition* on IM. The basic idea is that it partition IM into quadrants recursively if a quadrant is not coherent. The partition is called by the following two conditions:

1. the variance of a sub-quadrant $var(\Omega_i)$ is greater than $t * var(\Omega)$;
2. the mean of a sub-quadrant $mean(\Omega_i)$ is greater than $mean(\Omega)$.

where var and $mean$ calculate the variance and mean of flow direction and magnitude, respectively. t is a threshold. Ω is the entire IM. The condition 1 is for detecting the motion discontinuities including the boundaries of outliers and noise, where the flow value changes violently. The condition 2 is for detecting outliers' body, where the outliers' flow differs from inlier's flow because inlier has been almost canceled by two potentials in IM. Partition performs until no region can be split further. The smallest regions represent outliers, motion discontinuities and noise. The rest larger regions represent the coherent inlier, and will be involved in inlier approximation in next section.

The threshold t determines how the Quad-Tree partitions IM. Since local deformations usually vary on different IMs, it is rather difficult to find the best partition using one threshold. We, therefore, define a set of thresholds in a descending order, and introduce a progressive Quad-Tree partition. The t is initially set to 1, and reduces with a step 0.05 for next partition. The procedure stops when the difference between the current Quad-Tree QT' and the previous one QT is less than a convergence value ε. The updated QT' is used for labeling incoherence in IM. The pseudo-code is

Data: IM
Result: Labeled incoherence
$t = 1$; $QT = 0$;
Split(Ω);
while $|QT' - QT| > \varepsilon$ **do**
  for *each quadrant* Ω_i **do**
    if $var(\Omega_i) > t * var(\Omega) \parallel mean(\Omega_i) > mean(\Omega)$ **then**
      Split(Ω_i);
      Go to for loop;
    end
  end
  Update QT' and QT;
  $t = t - 0.05$;
end
Mark all smallest blocks as incoherence;

Algorithm 1: Progressive Quad-Tree Partition

Figure 1.(f) shows the result of progressive Quad-Tree partition which effectively labels outliers and depth discontinuities on IM.

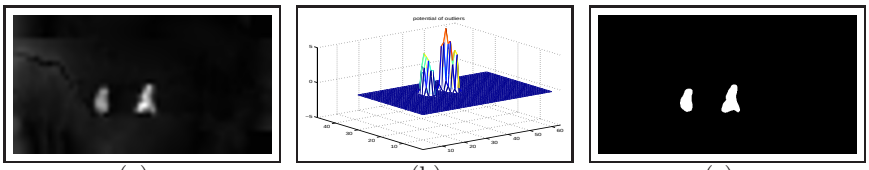
2.4 Inlier estimation

Multi-parametric models had been conducted to recover the inlier [12]. They were designed for camera motions ranging from simple translation to complex perspective transformation. But, the limitation is that a prior knowledge of motion structure is required to select an appropriate model. Nevertheless, this prior knowledge is not always available in real data. By contrast, we employ a general solution, polynomial surface fitting, using d-order polynomial

$$\mathcal{P} = a_{d0}x^d + a_{0d}y^d + \cdots + a_{ij}x^i y^j + \cdots + a_{10}x + a_{01}y + a_{00}, \tag{15}$$

to estimate inlier $\mathcal{F}_{in}$ from two potentials E and $\overrightarrow{W}$. The advantage is that it requires no prior knowledge.

It is known that high order terms in Eq.(15) present the high frequency signals (incoherence: outliers, motion discontinuities and noise in this work). Thanks to IM and algorithm 1, the incoherence has been labeled and rejected in process afterwards. As explained in Eq.(6) and Eq.(11), our object is to find low order polynomial $\mathcal{P}'$ which expresses the coherent inlier. The inlier estimation, eventually, can be performed by sampling the rest flows on two potentials E and $\overrightarrow{W}$. Since outliers and noise are completely excluded, surface fitting utilizes nearby samples to approximate the inlier. Thus, the result is rather coherent with the camera motion. For motion discontinuities, surface fitting interpolates the samples at both sides of discontinuity to approximate the violent flow change. So, the result also presents the trend of rapid flow change. But the gradient of the change becomes less than the one on original $\mathcal{OF}$. In our work, a polynomial of $d = 5$ is employed to produce coherent and accurate potentials E' and $\overrightarrow{W}'$



(a) (b) (c)

Fig. 2. (a) The outliers map from Eq.(17). The flow discontinuities have become very weak. (b) The potential of outliers. (c) The labeled outliers.

using the samples after rejecting incoherence. The final inlier is estimated by a linear combination where α, β have been determined in section 2.3,

$$\mathcal{F}'_{in} = \alpha \nabla E' + \beta \nabla \times \overrightarrow{W}'. \tag{16}$$

Figure 1.(g) shows the estimated inlier. It demonstrates that our method approximates the inlier rather coherent, and the outliers have been excluded effectively.

3 Outliers detection

With estimated inlier, outliers $\mathcal{F}_{out}$ can be detected directly by subtracting the $\mathcal{F}'_{in}$ from the original optical flow $\mathcal{OF}$,

$$\mathcal{F}_{out} \cong \mathcal{OF} - \mathcal{F}'_{in}. \tag{17}$$

Please see Fig.2.(a). However, low order polynomial surface fitting has a defect that it better fits the data with dense samples but goes wild at the edges of the original domain Ω due to lack of adequate samples. To reduce the error, the result is filtered by the mean-curvature of the original potential. The final segmentation is obtained subsequently by assigning binary labels on the true outliers. We will use the segmentation result to evaluate the performance of proposed method in experiments. Figure 2.(c) and (b) show the detected outliers and their potentials, respectively. Figure 1.(h) shows the 3D segmentation result.

4 Experiments

The performance of proposed method is evaluated on four benchmark datasets: Hopkins [13], Berkeley Motion Segmentation [5], Complex Background [4], and SegTrack [14]. The Hopkins dataset contains video sequences along with the features extracted and tracked in all the frames, which has three categories: checkerboard, car, and people sequences. Since checkerboard sequences do not correspond to natural scenes. We just use one sequence (1R2TCR) to show the effectiveness of our method in dealing with cameras rotation. The Berkeley dataset is derived from the Hopkins dataset, which consists of 26 moving camera videos: car, people, and Marple sequences. This dataset has full pixel-level annotations on multiple objects for a few frames sampled throughout the video. Since Marple

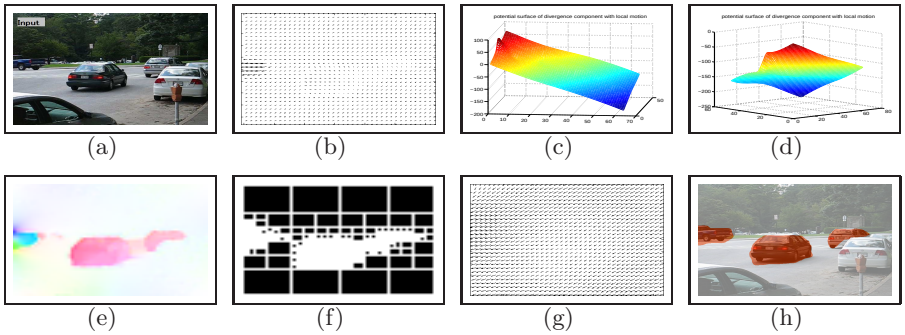


Fig. 3. (a) Cars2 sequence. (b) The original optical flow. Camera motion is translation. (c), (d) Two potentials of E and W. (e) IM. (f) Progressive Quad-Tree Partition on IM. (g) The estimated inlier. (h) Segmentation result.

sequences mainly contain static scenes or the objects are static, it is little challenging for our method. Thus, they are not involved in the experiments. Complex Background and SegTrack datasets contain extremely challenging scenes, where the background motion is much more complex than other datasets. They also provide full pixel-level annotations on multiple objects at each frame within each video. These two datasets are employed to highlight the advantage of our method.

We first illustrate the performance of the proposed method on three sequences that consist of camera translation, rotation and zoom in/out, respectively. Then we compare our method with state-of-the-art [4] on all four datasets. Optical flow is calculated using Brox's method [15] and optimized by [16].

4.1 Performance on three typical sequences

We demonstrate the performance of our method on three typical sequences: *cars2, 1R2TCR*, and *drive* that involve varied camera motions.

[1] Cars2 Sequence: This sequence is from the Berkeley dataset. Three cars are translating in the scene. The camera is translating too, please see Fig.3.

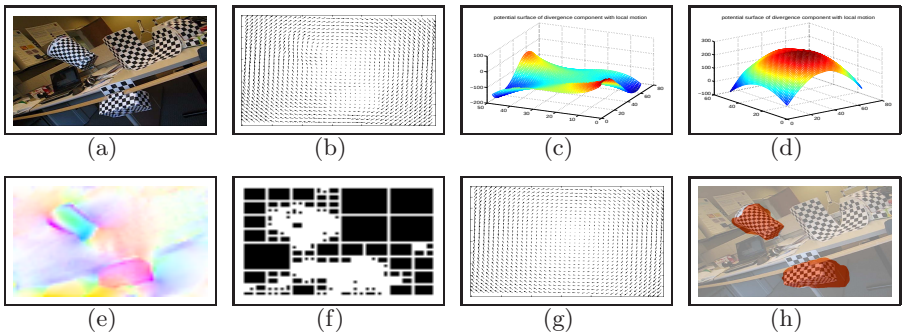
[2] 1R2TCR-Checkerboard Sequence: This data is from the Hopkins dataset. The basket is rotating, and the box is translating from left to the right. The camera is rotating, please see Fig.4.

[3] Drive sequence: This sequence is from the Complex Background dataset. A car is turning to left at the corner. The camera is zooming in, please see Fig.5.

For more examples, please refer the supplemental material.

4.2 Comparison with state-of-the-art

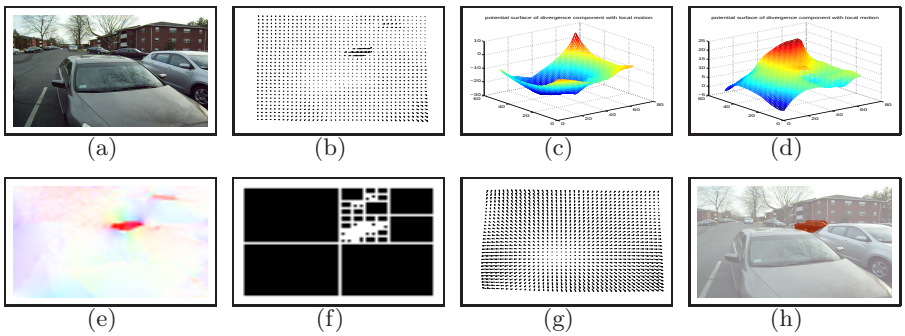
The proposed method is compared with the latest dense motion field based approach [4], which present two versions: (1) FOF, which uses optical flow information only, and (2) FOF+color+prior, which combines optical flow, color appearance and a prior model to improve the accuracy. We reported the F-measure of ours, FOF and FOF+color+prior presented in their paper in Table



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 4. 1R2TCR sequence. Please refer to Fig.3 for the description of subfigures.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 5. Drive sequence. Please refer to Fig.3 for the description of subfigures.

1. F-measure is employed because it considers both the precision P_r and the recall R_c of the test to compute the score as [17]:

$$F = \frac{2 \times R_c \times P_r}{R_c + P_r}.$$

Table 1 shows that our method outperform FOF and FOF+color+prior on almost all videos: it raises the F-measure by 10% – 30% on *Cars* 2, 3, 4, 7, and *People* 1 sequence in the Berkely dataset; around 10% on the *drive, parking*, and *store* sequences in the Complex Background dataset; and more than 20% on the *parachutte* and *monkeydog* sequences in the SegTrack dataset. This result is quite appealing even on videos containing extremely challenging scenes, such as the ones with occlusions, complex backgrounds, and noises. A few other results are shown in Fig.6, where the last column is the ground truth. In most cases, our segmentation agrees with the ground truth more than existing methods.

Other relevant works are Ochs *et al.*[6], Elqursh and Elgammal[7] and Kwak *et al.*[8]. These methods analysis trajectories using multiple frames, and some also need special initialization at the first frame. Thus, they are not directly comparable to inlier and outliers accuracy measure. Both our method and FOF are based on optical flow, and a frame-to-frame method requiring neither initial-

Table 1. F-measures of FOF method [4], and ours.

Sequences	FOF	FOF+color	Our	Sequences	FOF	FOF+color	Our
Cars1	47.81	50.84	**76.38**	drive	30.13	61.80	**83.03**
Cars2	46.37	56.60	**83.23**	forest	19.48	31.44	**35.81**
Cars3	67.18	73.57	**87.47**	parking	43.47	73.19	**83.57**
Cars4	38.51	47.96	**84.52**	store	28.46	70.74	**80.10**
Cars5	64.85	70.94	**84.92**	traffic	66.08	71.24	**71.77**
Cars6	78.09	84.34	**85.81**				
Cars7	37.63	42.92	**86.10**	birdfall2	68.68	75.69	**76.23**
Cars8	87.13	87.61	**90.78**	girl	75.73	**81.95**	78.06
Cars9	68.99	66.38	**77.52**	parachute	51.49	54.36	**86.72**
Cars10	53.98	50.84	**54.93**	cheetah	12.68	22.31	**55.77**
People1	56.76	69.53	**80.14**	penguin	14.74	20.71	**21.71**
People2	85.35	88.40	**89.91**	monkeydog	10.79	18.62	**45.45**

ization nor prior knowledge. Therefore, we only compare our method with FOF in this paper only.

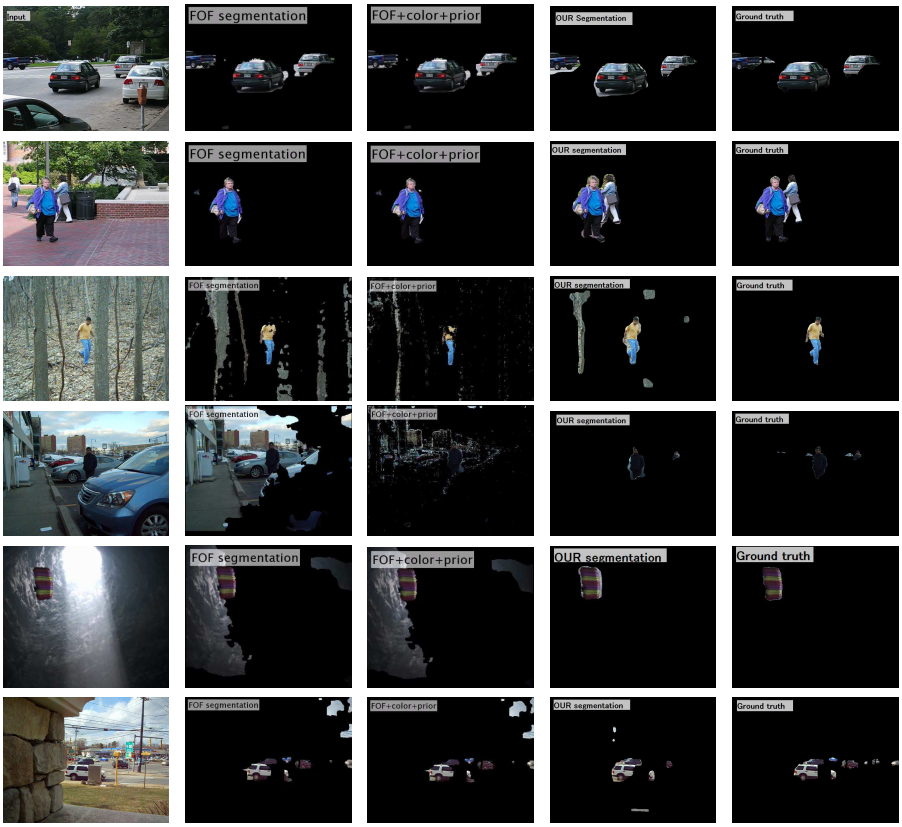
Although, the proposed method achieves inspiring performance, extensive experiences shows it may fail in the following cases:

- Motions of moving objects are very weak comparing with the camera motion. In this case, the outliers are more likely to be decomposed by HHD because they are pretty coherent with the inlier, and can not appear in IM. Such as the *cars* 1, 9 and 10 sequences and *forest* contain some objects' motions which are very small.
- A few isolated static objects stand alone in a texture-free background while camera is moving. In this case, our method may mistake these isolated static objects as moving objects.
- The outlier is greater than inlier. In this case, HHD fails to decompose the inlier because of the global minimization.

We have to point out that the accuracy of the optical flow affects the performance of our method as well. The *girl* sequence shows such an example. It captures a fast running girl in the sports yard. Some frames are blurred terribly, and have severe noisy optical flows. In this case, only optical flow is not sufficient. That's why FOF+color+prior utilizes additional information (color appearance) and prior models to improve the performance. In addition, both methods appear less accurate on the three sequences (*cheetah, penguin, monkeydog*) in the SegTrack dataset. The reason is they have multiple moving objects, but the ground truth intended for tracking one primary object as the foreground, causing all methods appear less accurate.

5 Conclusions

We have presented a general framework for 3D motion segmentation on a wide bank of moving camera videos. This framework solved two problems: 1) the



(a) (b) (c) (d) (e)

Fig. 6. Segmentation results of FOF, FOF-color and ours on challenging scenes: (a) input sequences, from top to bottom: cars2, people2, forest, store, parachute, traffic, (b) FOF, (c) FOF+color+prior, (d) our segmentation, (e) ground-truth segmentation.

amended HHD can handle the coherent inlier of all three primary motion flows (translation, rotation, and radial flow), 2) the proposed Incoherence Map and progressive Quad-Tree precisely label the outliers, motion discontinuities and noise. The afterward inlier estimation is achieved by approximating low order polynomials using the rest samples on two potentials in HHD. This compensates the depth discontinuity in the 3D motion. We have evaluated our approach on four benchmark datasets. Extensive experiments showed a rather comparable performance than state-of-the-art. In the future work, more coherent information (e.g. colors, the direction only) might further help our method for segmentation.

Acknowledgement. This work is supported by: Japan Society for the Promotion of Science, Scientific Research KAKENHI for Grant-in-Aid for Young Scientists (ID:25730113).

References

1. Chen, Y.M., Bajic, I.V.: Motion vector outlier rejection cascade for global motion estimation. IEEE Signal Processing Letters **17** (2010) 197–200
2. Chen, Y.M., Bajic, I.V.: A joint approach to global motion estimation and motion segmentation from a coarsely sampled motion vector field. IEEE Transactions on Circuits Sysstem and Video Technology **21** (2011) 1316–1328
3. Qian, C., Bajic, I.V.: Global motion estimation under translation-zoom ambiguity. In: Proc. IEEE PacRim. (2013) 46–51
4. Narayana, M., Hanson, A., Learned-Miller, E.: Coherent motion segmentation in moving camera videos using optical flow orientations. In: ICCV. (2013)
5. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: ECCV. (2010)
6. Ochs, P., Malik, J., Brox, T.: Segmentation of moving objects by long term video analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2014)
7. Elqursh, A., Elgammal, A.: Online moving camera background subtraction. In: ECCV. (2012) 228–241
8. 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) 2174–2181
9. Irani, M., Rousso, B., Peleg, S.: Recovery of ego-motion using image stabilization. In: CVPR. (1994)
10. Helmholtz, H.: On integrals of the hydrodynamical equations, which express vortex-motion. Philosophical Magazine and J. Science **33** (1867) 485–512
11. Bhatia, H., Norgard, G., Pascucci, V., Bremer, P.T.: The helmholtz-hodge decomposition - a survey. IEEE Transactions on Visualization and Computer Graphics (TVCG) **19** (2013) 1386–1404
12. Su, Y., Sun, M.T., Hsu, V.: Global motion estimation from coarsely sampled motion vector field and the applications. IEEE Transactions on Circuits Sysstem and Video Technology. **15** (2005) 232–242
13. Tron, R., Vidal, R.: A benchmark for the comparison of 3D motion segmentation algorithms. In: CVPR. (2007)
14. Tsai, D., Flagg, M., M.Rehg, J.: Motion coherent tracking with multi-label mrf optimization. In: BMVC. (2010)
15. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: ECCV. (2004) 25–36
16. Liang, X., McOwan, P., Johnston, A.: A biologically inspired framework for spatial and spectral velocity estimations. Journal of the Optical Society of America A **28** (2011) 713–723
17. Dembczynski, K., Jachnik, A., Kotlowski, W., Waegeman, W., Hullermeier, E.: Optimizing the F-measure in multi-label classication: Plug-in rule approach versus structured loss minimization. In: International Conference on Machine Learning (ICML). (2013) 1130–1138