Motion Segmentation by Exploiting Complementary Geometric Models

Xun Xu
National University of Singapore
elexuxu@nus.edu.sg

Loong Fah Cheong
National University of Singapore
eleclf@nus.edu.sg

Zhuwen Li
Intel Labs
li.zhuwen@intel.com

Abstract

Many real-world sequences cannot be conveniently categorized as general or degenerate; in such cases, imposing a false dichotomy in using the fundamental matrix or homography model for motion segmentation would lead to difficulty. Even when we are confronted with a general scene-motion, the fundamental matrix approach as a model for motion segmentation still suffers from several defects, which we discuss in this paper. The full potential of the fundamental matrix approach could only be realized if we judiciously harness information from the simpler homography model. From these considerations, we propose a multi-view spectral clustering framework that synergistically combines multiple models together. We show that the performance can be substantially improved in this way. We perform extensive testing on existing motion segmentation datasets, achieving state-of-the-art performance on all of them; we also put forth a more realistic and challenging dataset adapted from the KITTI benchmark, containing real-world effects such as strong perspectives and strong forward translations not seen in the traditional datasets.

1. Introduction

Various geometric models have been used in the motion segmentation problem to model the different types of cameras, scenes, and motions. In this problem as commonly set forth, the underlying models are generally regarded as applicable under different scenarios and these scenarios do not overlap. For instance, when the underlying motion is a general motion, fundamental matrix is used to model the epipolar geometry [16, 23], and when scene-motion is degenerate like a planar scene or a pure rotation, homography is preferred [6, 18]. However, the real world scene-motions are in fact not so conveniently divided: they are more typified by near-degenerate scenarios such as a scene that is almost but not quite planar, or a motion that is rotation-dominant but with a non-vanishing translation. In such cases, imposing a false dichotomy in deciding an appropriate model would pose difficulty for subsequent subspace separation. For instance, it is well-known [11, 27, 30] in the case of a scene with dominant-plane, it is easy to find inliers belonging to the degenerate configuration (the plane), but the precision of the resulting fundamental matrix is likely to be very low. Most of the inliers outside the degenerate configuration will be lost, and often the erroneous fundamental matrix will pick up outliers (e.g. from other motion groups). Since this is not a purely planar scene, using homography in a naive manner might fail to group all the inliers together too, resulting in over-segmentation of the subspaces.

It is also not hard to establish—from a glance of the motion segmentation literature—that of the various models, the fundamental matrix model is generally eschewed, due to the lack of perspective effects in the Hopkins155 benchmark [31]. However, it is never clearly articulated if the numerical difficulties arising from degeneracies in such approach present insuperable obstacles. And no one has put his/her finger on the exact manner how the resulting affinity matrix is ill-suited for subspace clustering: is it solely due to the degeneracies or are there other factors? Considering that in many real-world applications say, autonomous driving, perspective effects are not uncommon, it surely follows that we should come to a better understanding of the suitability of fundamental matrix (or for that matter, the homography model) as a geometric model for motion segmentation. This, we contend, is far from being the case. For instance, does it follow that if we use the fundamental matrix for wide field-of-view scenes, like those found in the KITTI benchmark [9], we will get better performance than those using homography? We have in fact as yet no reason to believe that this will be the case, judging by the way how the various algorithms based on affine model still outperform those based on fundamental matrix in individual Hopkins sequences that have larger perspectives (though admittedly still moderate). Indeed, from the results we obtained on the KITTI sequences that we adapted for testing motion segmentation in real-world scenarios, the superiority of the homography-based methods is again observed. Thus, one might naturally ask what factors other than degeneracies are hurting the fundamental matrix approach? And why is the homography matrix approach holding its own in wide per-

spective scenes, when it possesses none of the geometrical exactness of the fundamental matrix?

In the remainder of this section, we will briefly investigate the suitability of homography and fundamental matrices ($\mathbf{H}$ and $\mathbf{F}$ respectively) as a geometric model for motion segmentation. We shall henceforth denote the affinity matrices generated by $\mathbf{H}$ and $\mathbf{F}$ as $\mathbf{K_H}$ and $\mathbf{K_F}$ respectively.

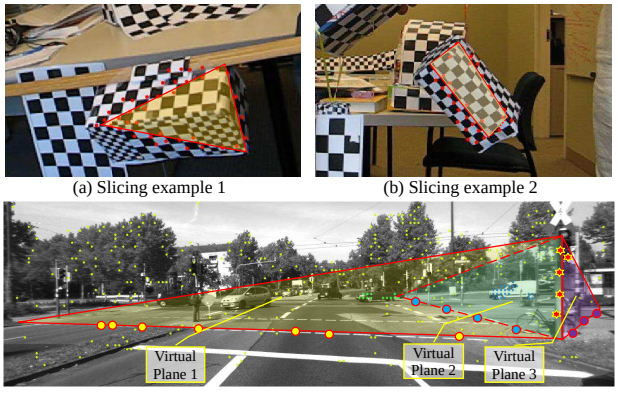
1.1. Success roadmap of H

The preceding paragraphs have already alluded to the fact that the affinity matrix $\mathbf{K_H}$ may not exhibit high intra-cluster cohesion (due to lack of strong affinity between different planes of the same rigid motion), and thus might lead one to be skeptical of its adequacy for the purpose of motion segmentation. In the Hopkins155 dataset, this is not an overriding concern since most of the sequences have small field-of-view and perhaps the scene is sufficiently far away to be well approximated by a plane; these approximations are seemingly borne out by the good empirical results obtained by a wide variety of approaches based on affine subspace or homography matrix. The recent homography-based method [18] boasts state-of-the-art performance with a mean error of 0.83%. The low error attained is noteworthy given that there are actually some Hopkins sequences with non-negligible perspective effects; we feel that this phenomenon warrants a better explanation than the reasoning offered so far.

The success can be attributed to the many planar slices induced by the homography hypothesizing process; these are not necessarily actual physical planes in the scenes (see the slices in Fig. 1 (a-b)) but as long as these virtual planes belong to the same rigid motion, it is evident that they can be fitted with a homography. Such slicings of the scene create strong connections between points across multiple real planar surfaces and result in a much less over-segmented affinity matrix $\mathbf{K_H}$. If the scene contains only compact objects or piecewise smooth structures, then such connectivity created is sufficient to bind the various surfaces of a rigid motion together. However, in the real world sequences, when the above conditions are not satisfied, we suspect that this may not be adequate. Fig. 1(c) illustrates a background comprising an elongated object (a traffic light) and the marking on the road. It is clear that in this case, while one can form virtual planar slices as before, the resulting connectivity is much lower (most if not all of the slices cannot connect large segments of both these elements simultaneously, unlike those in Fig. 1 (a-b)).

1.2. Problems with F

Besides the degeneracy issues that are well-known from the classical structure from motion literature, we suggest that another root problem with the fundamental matrix approach for the motion segmentation problem lies precisely



(a) Slicing example 1 (b) Slicing example 2

(c) Slicing example in real world

Figure 1: Illustration of slicing effect of homography. (a-b) Red dots indicate inlier points of a hyopthesis. All points lie on a virtual plane (a slice of the cube) highlighted in yellow. (c) Virtual planes are highlighted as triangles with points in the same color as inliers.

in the fact that it is an all-encompassing model that captures all types of scene-motion configurations. The risk of such a complicated model for the subsequent clustering and model-selection task is not difficult to surmise. The richness of characterization renders it likely to capture any correlation between different rigid motions. Therefore it is more likely to cause overlapping the subspaces of different rigid motions than simpler models, e.g. homography. However, we find ourselves asking, is it not possible that $\mathbf{F}$ also offers the greatest scope for forming the best correct view, given that it starts with a geometrically correct model that the homography model can hardly be, and the former must have thus captured much of what is correct? It perhaps just requires some nudge in the correct direction for us to reclaim the performance that ought be had for $\mathbf{K_F}$. From this standpoint, even when we are confronted with a general scene-motion with no degeneracies, there is still an important reason for keeping the homography model—to midwife the unborn view of $\mathbf{K_F}$.

1.3. Proposed solution

We have been at pains to point out that many real-world sequences cannot be classified into neat categories such as general or degenerate scene-motions and thus cannot be adequately addressed by any single model such as $\mathbf{H}$ or $\mathbf{F}$. We have also discussed the defects of the fundamental matrix approach and conjectured that even though the resulting $\mathbf{K_F}$ may not have committed itself to any definite view on the potential clusters, its full potential could perhaps be realized if we judiciously harness information from a simpler model such as $\mathbf{H}$. From these considerations, we propose a multi-view[1] spectral clustering framework that synergisti-

[1]Note that this "view" here refers to the view from the standpoint of a model and should not be confused with the camera viewpoint.

cally combines these multiple models together. As there is no definite consensus on how best to combine several views together for spectral clustering, we evaluate a few extant fusion schemes. By doing so, we make sure that our findings are not an artifact of a particular fusion scheme. As we will show later, the performance of the fundamental matrix approach can be substantially raised using the improved $\mathbf{K_F}$. We hasten to add that one should not over-claim the potential gains of this fundamental matrix approach. When the scene contains substantial amount of degeneracies, as real scenes are apt to be, it is always better to rely on the combined view for the best performance. That is, one should seek a common spectral embedding that takes into account both the improved $\mathbf{K_F}$ and the improved $\mathbf{K_H}$.

To summarize, the contributions of our paper are as follows. First, we contribute to an understanding of the strengths and drawbacks of homography and fundamental matrices as a geometric model for motion segmentation. We then propose using affinity matrix fusion as a means of dealing with real-world effects that are often difficult to model with a pure homography or fundamental matrix. Finally, we perform extensive testing on existing motion segmentation datasets, achieving state-of-the-art performance on all of them; we also put forth a more realistic and challenging dataset adapted from the KITTI benchmark, containing real-world effects such as strong perspectives and strong forward translations not seen in the traditional datasets.

2. Related Work

A long line of works have studied the motion segmentation problem from different perspectives. They can be divided into two major groups: those based on a hypothesis-and-test paradigm and those that are more analytic rather than hypothesis-driven. Into the latter camp falls a wide variety of approaches, including factorization [1, 5, 8, 12, 29], algebraic method [26, 32, 33, 34], affinity matrix [20, 39], including those constructed from sparse [7] and low-rank [24] representations. They typically assume that the input is made up of a union of motions of specific types, with only a few works [10, 26] that can handle mixed types of motions. These analytic approaches are rightly praised for their elegance but become awkward in dealing with real world signals that are often drawn from mixed multiple manifolds. In contrast, works in the former category, being hypothesis-driven, are naturally more suited to handling mixed models. This is exemplified in the earlier works such as [27, 30] which explicitly decide on whether $\mathbf{F}$ or $\mathbf{H}$ is better suited as a motion model in the face of possibly degenerate scene-motion configuration, but these works are applied to cases where the background is by far the most dominant group in the scene. Subsequent hypothesis-and-test methods [3, 4, 21] dealing with the realistic Hopkins155 [31] sequences almost as a rule ignore the more complex fundamental matrix (or equivalently the perspective projection model) altogether, possible reasons being the computational complexity issues posed by the outliers under the fundamental matrix model and/or the lack of perspective effects in the Hopkins sequences. Thus, these later works do not concern themselves with the problem of dealing with mixed types of models. Our approach differs from the above works in that not only do we allow for mixed types of models, we also do not impose a dichotomous decision on what is an appropriate model.

Spectral clustering has been an attractive tool for clustering data [35]. Under this framework there are roughly two genres. The first kind discovers an optimal combination to aggregate multiple affinity matrices (kernels) for spectral clustering [14, 19, 36]. However such combination is often non-trivial to discover. Alternatively, studies have been carried out on discovering a consensus on multiple kernels. In particular, the co-regularization scheme [17] was proposed to force data from different views to be close to each other in the embedding space for clustering. Few if any of the existing approaches can guarantee superiority to the simple approach—kernel addition. In this work, we start our evaluation with this simplest baseline and then reveal its relation with the co-regularization schemes. We also evaluate a custom-built version incorporating a subset constraint that preserves the true hierarchical structure of the affinity matrices induced by different geometric models.

3. Methodology

In this section, we first describe the geometric models used for motion segmentation and their hypothesis formation process. We then explain how the affinities between feature points are encapsulated in the ORK kernel[18]. Finally, we explain the extension from single-view to multi-view clustering. In particular, we elaborate the relation between kernel addition and co-regularization for generic multi-kernel clustering, and we describe how the geometric relation that exists between models can be used to formulate a custom-made subset constrained multi-view clustering.

3.1. Geometric Model Hypothesis

Denote the observations of tracked points throughout F frames as $\{\mathbf{x}_i\}_{f=1\cdots F}$. We then randomly sample a minimal number of p such points visible in a pair of frames and use them to fit a hypothesis of the model. The models tested include the fundamental matrix $\mathbf{F}$, homography $\mathbf{H}$, as well as the affine matrix $\mathbf{A}$. The reason for including the affine matrix model is because many existing datasets contain sequences with very weak perspective so this simpler model might be numerically more stable. For the three models $\mathbf{F}$, $\mathbf{H}$, and $\mathbf{A}$, the respective values for p are 8, 4, and 3. The parameters of the model are estimated via linear algorithms

[13] and $500 \times F$ hypotheses are sampled for each type of geometric model.

3.2. Affinity Captured as Ordered Residual Kernel

Given multiple hypotheses $\{\mathbf{Y}_k\}_{k=1\cdots K}$ generated from a particular model (affine, homography or fundamental matrix), we first compute for each data point the residual to all these hypotheses $\{d(\mathbf{x}_i, \mathbf{Y}_k)\}_{k=1\cdots K}$ in terms of their Sampson errors [13]. The affinity between two features is captured in the correlation of preference for these hypotheses. Specifically, we can define the correlation in terms of the co-occurrence of points among all hypotheses. That is, if we define the indicator of point $\mathbf{x}_i$ being the inlier of all hypotheses $\{\mathbf{Y}_k\}$ as $\mathbf{o}_i \in \{0,1\}^K$, then the co-occurrence between two points is written as $k_{ij} = \mathbf{o}_i^\top \mathbf{o}_j$. However, the threshold τ needed to determine when a data is an inlier (i.e. $\mathbf{o}_i = \mathbb{1}(d(\mathbf{x}_i, \mathbf{H}_k) < \tau)$) is not easy to set, due to the potentially disparate range of motions present in different sequences. The ordered residual kernel (ORK) [3, 18] was proposed to deal with this issue. Instead of fixing a threshold, the ORK sorts the residual in ascending order $\{\hat{d}_{i1} \quad \hat{d}_{i2} \cdots \hat{d}_{iK}\}$ where $\forall k : \hat{d}_{ik} <= \hat{d}_{ik+1}$. An adaptive threshold is then selected as the top h-th residual, i.e. $\tau_i = \hat{d}_{ih}$. The ORK kernel is also known to be resilient to serious sampling imbalance, an important advantage in real-world scenes where background is usually very large. Therefore, we adopt the ORK kernel to encapsulate the affinities between feature points. After constructing the affinity matrix, we normalize the affinities by dividing all k_{ij} entries by the number of frames where both feature points i and j are visible. This step removes the weighting balance caused by incomplete trajectories. Finally, as is customary in motion segmentation works, we subject the affinity matrix to a sparsification step; we use the ϵ-neighborhood scheme of [18] for this purpose.

3.3. Spectral Clustering for Motion Segmentation

We are now ready to use spectral clustering to recover the clusters. We first review the single view spectral clustering problem and then extend it to multi-view clustering.

3.3.1 Single-View Spectral Clustering

Given the single affinity matrix $\mathbf{K}$, the normalized Laplacian $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-0.5}\mathbf{K}\mathbf{D}^{-0.5}$ is first computed, where $\mathbf{D}$ is the degree matrix. The following objective is then set up to eigendecompose $\mathbf{L}$:

$$\min_{\mathbf{U}} tr\left(\mathbf{U}^\top \mathbf{L}\mathbf{U}\right), \quad s.t. \mathbf{U}\mathbf{U}^\top = \mathbf{I} \tag{1}$$

where $tr(\cdot)$ is the trace operator. The spectral embedding $\mathbf{U} \in \mathbb{R}^{N\times M}$ can be efficiently solved and then treated as a new feature representation of the original points. A separate K-means step is then fed with the first M dimensions of the normalized $\mathbf{U}$ for grouping points into M motions.

3.3.2 Multi-View Spectral Clustering

With multiple views provided by the different types of motion models, we have now at our disposal multiple affinity matrices. We explore two generic and one custom-made multi-view spectral clustering schemes to fuse the multiple sources of information together for clustering.

Kernel Addition A naive way to fuse information from heterogeneous sources for clustering is by kernel addition [17]. Given affinity matrices induced by heterogeneous sources $\{\mathbf{K}_v\}_{v=1\cdots V}$, kernel addition yields a fused kernel by summing up each individual kernel $\mathbf{K} = \sum_v \mathbf{K}_v$. With the corresponding Laplacian matrices written as $\mathbf{L}_v = \mathbf{I} - \mathbf{D}_v^{-0.5}\mathbf{K}_v\mathbf{D}_v^{-0.5}$, the objective for kernel addition can be written as,

$$\begin{aligned} &\min_{\mathbf{U}} tr(\mathbf{U}^\top \sum_v \mathbf{L}_v \mathbf{U}), \quad s.t. \mathbf{U}^\top\mathbf{U} = \mathbf{I} \\ \Rightarrow &\min_{\{\mathbf{U}_v\}} \sum_v tr(\mathbf{U}_v^\top \mathbf{L}_v \mathbf{U}_v), \quad s.t. \mathbf{U}_v^\top \mathbf{U}_v = \mathbf{I}, \\ &\forall v, w \in \{1, \cdots V\} : \mathbf{U}_v = \mathbf{U}_w \end{aligned} \tag{2}$$

We notice the kernel addition strategy is equivalent to discovering a common spectral embedding $\mathbf{U}$ among all views. This requirement of having a single consensus embedding can be too strong.

Co-Regularization Instead of demanding a common embedding, another solution is to include an additional regularization term in the objective function to encourage pairwise consensus between any two spectral embeddings $\mathbf{U}_v$ and $\mathbf{U}_w$. This has been studied by [17] who introduced a co-regularization term $tr\left(\mathbf{U}_v\mathbf{U}_v^\top \mathbf{U}_w\mathbf{U}_w^\top\right)$. This trace term returns high value if the new kernel matrix in the spectral embedding space $\mathbf{U}_v\mathbf{U}_v^\top$ and $\mathbf{U}_w\mathbf{U}_w^\top$ are similar to each other and vice versa. Incorporating the co-regularization term, we obtain the following objective:

$$\begin{aligned} &\min_{\{\mathbf{U}_v\}} \sum_v tr(\mathbf{U}_v^\top \mathbf{L}_v \mathbf{U}_v) - \lambda \sum_v \sum_{w, w\neq v} tr(\mathbf{U}_v\mathbf{U}_v^\top \mathbf{U}_w\mathbf{U}_w^\top), \\ &s.t. \mathbf{U}_v^\top \mathbf{U}_v = \mathbf{I} \end{aligned} \tag{3}$$

We can interpret the co-regularization scheme as a relaxed version of kernel addition. By increasing the penalty coefficient λ, the co-regularization scheme will approach kernel addition as all embeddings are forced to approach each other. This model is termed as pairwise co-regularization by [17] as the co-regularization term comprises of all pairs of spectral embeddings. The co-regularization model can be efficiently solved by initializing each view $\mathbf{U}_v$ separately in the same way as single-view spectral clustering. Then we recursively update each view

with all other views fixed. When solving a single view, the problem becomes a standard eigendecomposition problem. After convergence, we can concatenate the new spectral embedding of all views to produce an extended feature for the K-means step.

3.3.3 Subset Constrained Multi-View Spectral Clustering

The above two multi-view spectral clustering schemes are generic fusion methods that do not exploit any relation that might exist between the different views. In the specific case of motion segmentation, we know that for any $\mathbf{H}$ between two views, we can always define a family of $\mathbf{F} = [\mathbf{e}]_x \times \mathbf{H}$ parameterized by a vector $\mathbf{e}$, where $[\mathbf{e}]_x$ denotes the skew-symmetric matrix of $\mathbf{e}$ [13]. This means a pair of points that are the inliers of a homography should always be the inliers of a certain fundamental matrix. Conversely, if a pair of points are not the inliers of any $\mathbf{F}$, there is no homography which could take both points as inliers[2]. Generally speaking, we should expect that if $\mathbf{K_A}$, $\mathbf{K_H}$, and $\mathbf{K_F}$ are ideal binary affinity matrices, then $\mathbf{K_A} \leq \mathbf{K_H} \leq \mathbf{K_F}$. We term this hierarchical relationship the subset constraint. Imposing this constraint will help to further denoise or repair the affinity matrices. We cast this problem as a constrained clustering problem (adapted from [37]):

$$\begin{aligned} &\min_{\{\mathbf{U}_v\}} \sum_v tr\left(\mathbf{U}_v^\top \mathbf{L}_v \mathbf{U}_v\right) - \gamma tr(\mathbf{U}_v^\top \mathbf{Q}_v \mathbf{U}_v), \\ &s.t. \mathbf{U}_v^\top \mathbf{U}_v = \mathbf{I}, \quad \mathbf{Q}_v \in \{-1, 0, 1\}^{N \times N} \end{aligned} \tag{4}$$

where the matrix $\mathbf{Q}_v$ provides the subset constraint for the v-th view. For $q_{ij} = 1$, the constraint encourages a high inner product $\mathbf{u}_{vi}^\top \mathbf{u}_{vj}$ where $\mathbf{u}_{vi}$ indexes the i-th column. This means points i and j are encouraged to fall into the same cluster. For $q_{ij} = -1$, the constraint encourages a different cluster assignment between i and j, and lastly, for $q_{ij} = 0$, there is no constraint. For any single view v, the constraints $\mathbf{Q}_v$ is imposed by other views. For example, solving view $\mathbf{H}$, the positive constraint q_{ij} is inherited from the result of $\mathbf{K}$; that is, if there is a link between points i and j from $\mathbf{K_A}$, then the (i, j) entry of $\mathbf{K_H}$ is encouraged to be 1. On the other hand, the negative constraints come from $\mathbf{F}$. One could solve this problem using an alternating minimization scheme, but the subset constraint matrix $\mathbf{Q}_v$ may flip their values from 1 to -1 and vice versa in each alternating step, posing significant difficulties for convergence.

Therefore, we relax $\mathbf{Q}_v$ to continuous values. Instead of utilizing the discretized results from other views, we use the affinity reconstructed from the spectral embedding $\hat{\mathbf{K}} = \mathbf{U}\mathbf{U}^\top$ to construct $\mathbf{Q}_v$ as detailed in Eq (5). We assume the three views are placed in the order of affine ($v = 1$), homography ($v = 2$) and fundamental matrix ($v = 3$). The final objective is then written as Eq (5).

$$\begin{aligned} &\min_{\{\mathbf{U}_v\}} \sum_v tr\left(\mathbf{U}_v^\top \mathbf{L}_v \mathbf{U}_v\right) - \gamma tr(\mathbf{U}_v^\top \mathbf{Q}_v \mathbf{U}_v), \quad s.t. \mathbf{U}_v^\top \mathbf{U}_v = \mathbf{I}, \\ &\mathbf{Q}_v = \begin{cases} \mathbb{1}\left(\hat{\mathbf{K}}_{v+1} < 0\right) \circ \hat{\mathbf{K}}_{v+1}, & v = 1 \\ \mathbb{1}\left(\hat{\mathbf{K}}_{v-1} > 0\right) \circ \hat{\mathbf{K}}_{v-1} + \mathbb{1}\left(\hat{\mathbf{K}}_{v+1} < 0\right) \circ \hat{\mathbf{K}}_{v+1}, & v = 2 \\ \mathbb{1}\left(\hat{\mathbf{K}}_{v-1} > 0\right) \circ \hat{\mathbf{K}}_{v-1}, & v = 3 \end{cases} \end{aligned} \tag{5}$$

where $\circ$ represents element-wise multiplication and $\mathbb{1}(\cdot)$ is the indicator function. The subset constraint means for view $\mathbf{A}$ ($v = 1$), only the negative constraint from $\mathbf{H}$ is applied, for view $\mathbf{H}$, both positive and negative constraints from $\mathbf{A}$ and $\mathbf{F}$ are applied respectively. The final problem can be solved by optimizing each view $\mathbf{U}_v$ in an alternating fashion. We summarize the whole procedure in Algorithm 1.

Algorithm 1: Subset Constrained Clustering

input : Kernel matrices $\{\mathbf{K}_v\}$, no. of motion M and γ
output: Rigid motion index s
Initialize Spectral Embedding
for $v \leftarrow 1$ **to** V **do**
  Compute Laplacian matrix $\mathbf{L}_v = \mathbf{I} - \mathbf{D}_v^{-0.5} \mathbf{K}_v \mathbf{D}_v^{-0.5}$;
  $\mathbf{U}_v \leftarrow$ first M eigenvectors of $\mathbf{L}_v$;
Subset Constrained Spectral Clustering
while *Not Converged* **do**
  for $v \leftarrow 1$ **to** V **do**
    Compute $\mathbf{Q}_v$ following Eq (5);
    Compute constrained Laplacian matrix $\tilde{\mathbf{L}}_v = \mathbf{L}_v - \gamma \mathbf{Q}_v$;
    $\mathbf{U}_v \leftarrow$ first M eigenvectors of $\tilde{\mathbf{L}}_v$;
K-means to return index
$\mathbf{U} \leftarrow$ `Concatenate(`$\mathbf{U}_1, \cdots \mathbf{U}_V$`)` ;
$s \leftarrow$ `K-means(`$\mathbf{U}$,M`)`

3.3.4 Convergence Analysis

For both co-regularization and subset constrained clustering, we note the objective is not guaranteed to be convex w.r.t. all views' embeddings. Nevertheless, we prove that the co-regularization model guarantees to converge to at least a local minimal. As we solve the problem in an alternating fashion, each step involves solving Eq (3) for v-th view with all other views fixed, i.e. $\min_{\mathbf{U}_v} tr\left(\mathbf{U}_v^\top \left(\mathbf{L}_v - \lambda \sum_{w, w \neq v} \mathbf{U}_w \mathbf{U}_w^\top\right) \mathbf{U}_v\right)$. Such problem can be efficiently solved by eigen decomposition regardless of the convexity of $\left(\mathbf{L}_v - \lambda \sum_{w, w \neq v} \mathbf{U}_w \mathbf{U}_w^\top\right)$. Therefore, solving all views iteratively results in monotonic

[2] We assume in the above two propositions that there are always enough points to fit an $\mathbf{F}$ if it exists.

decreasing cost until converging to a local minimal. The convergence for subset constrained clustering is, however, not guaranteed due to the constraint matrix $\mathbf{Q}_v$ changes at each iteration. Nevertheless, experiment results suggest a proper selection of λ, less than $1e-2$ renders the problem easy to converge.

4. Experiment

We carry out experiments on three extant motion segmentation benchmarks including the Hopkins155 [31], the Hopkins12 [26] for testing incomplete trajectories and MTPV62 [23] for testing stronger perspective effects. For all three datasets, we evaluate the performance in terms of classification error [31]. We also put forth a new dataset that is adapted from the KITTI benchmark [9], containing real-world effects such as strong perspectives and strong forward translations not seen in the traditional datasets.

4.1. Motion Segmentation on Existing Benchmarks

In this section, we extensively compare single-view and multi-view approaches on Hopkins155 benchmark [31]. Specifically, for single-view, we evaluate using affine, homography and fundamental matrix as the single geometric model. For multi-view motion segmentation, we evaluated Kernel Addition (KerAdd), Co-Regularization (CoReg) [17] and Subset Constrained Clustering (Subset). We fix the regularization parameter λ and γ at 10^{-2}. We also extensively compare with state-of-the-art approaches, including: ALC [26], GPCA [34], LSA [38], SSC [7], TPV [23], T-Linkage [25], S^3C [22], RSIM [15] and MSSC[18]. The results are presented in Table 1. For those algorithms which do not explicitly handle missing data, we recover the data matrix using Chen's matrix completion approach [2].

We make the following observations from the results. Firstly, with regards to the use of homography matrix as a single geometric model, our finding echoes the excellent results of earlier work such as MSSC [18]. In fact, the simpler affine model has an even lower error figures. Clearly, the stitching argument (via virtual slices) put forth in Section 1 for explaining the success of homography applies to the affine case too, in particular under weak perspective views. For the fundamental matrix as a model, the performance is slightly worse-off. The reasons are manifold: strong camera rotation, limited depth relief, and not least the subspace overlap between different rigid motions, to which this richer fundamental matrix model is particularly susceptible. Secondly, after fusing multiple kernels, we saw a boost in performance compared to single-view approaches, e.g. 0.36% error for kernel addition and 0.31% for subset constrained clustering on Hopkins155. Consistent boost in performance can be observed on Hopkins12 and MTPV62 as well. Usually, the fusion can produce the best of all performance regardless of the fusion scheme used. Even the simple kernel addition yields very competitive performance. This provides a strong option for real applications where parameter tuning is not desirable.

4.2. Motion Segmentation on KITTI Benchmark

The limitations of the Hopkins155 dataset are well-known: limited depth reliefs, dominant camera rotations, among others. Such a dataset cannot meet the requirements of a benchmark for investigating motion segmentation capability in-the-wild, in particular self-driving scenario where the camera platform is often performing large translation and the scene is considerably more complex. For this reason, we propose a new motion segmentation benchmark based on the KITTI dataset [9], the KITTI 3D Motion Segmentation Benchmark (KT3DMoSeg). We choose short video clips from the raw sequences of KITTI governed by three principles. Firstly, we wish to study sequences with more significant camera translation so camera mounted on moving cars are preferred. Secondly, we wish to investigate the impact of complex background structure, therefore, scene with strong perspective and rich clutter (in the structure sense) is selected. Lastly, we are interested in the interplay of multiple motions, so clips with more than 3 motions are also chosen, as long as these moving objects contain enough features for forming motion hypotheses. 22 short clips, each with 10-20 frames, are chosen for evaluation. We further extract dense trajectories from each sequence using [28] and prune out trajectories shorter than 5 frames. Illustration of sample frames with labelled ground-truth and further details about the dataset (such as the preprocessing of trajectories) are given in the supplementary material.

We fit hypotheses on all valid tracking points, i.e. dense background and the evaluation is carried out on subsampled background as introduced in the supplementary. The same set of evaluation as in the preceding subsection is carried out and the results are presented in Tab. 1. Both average and median classification errors are reported. The performances of the multi-view approaches are again consistently better than those of the single geometric model. Further evaluation on individual sequence is presented in Fig. 2 (a). To give some context to the performance figures, we use the "Prevalence" column to indicate the baseline solution of just assigning every feature as belonging to the prevalent group—the background. The overall performance of this baseline approach is 27.95% which is pretty strong compared to many existing approaches. For the more recent and hypothesis-driven approach like MSSC, although we do not have the codes for evaluation, we can get an idea of its performance in KT3DMoSeg by looking at the result of our single-view homography model, due to its essential similarity to MSSC. Clearly, the homography model is able to replicate its strong performance (11.45%) on this real-world dataset despite facing much stronger perspective ef-

Table 1: Motion segmentation results on Hopkins155, Hopkins12, MTPV62 and KT3DMoSeg datasets evaluated as classification error (%). *The best performing model (RPCA+ALC$_5$ is reported for ALC [26]). ** State-of-the-Art models' performances are reported for the sequences with correct number of motion. '−' cells indicate not reported or no public code is available.

Models	Hopkins155 [31]			Hopkins12 [26]		MTPV62 [23]**				KT3DMoSeg	
State-of-the-Art	2 Motion	3 Motion	All	Average	Median	Perspective 9 clips	Missing Data 12 clips	Hopkins 50 clips	All 62 clips	Average	Median
LSA [38]	4.23	7.02	4.86	-	-	-	-	-	-	38.30	38.58
GPCA [34]	4.59	28.66	10.02	-	-	40.83	28.77	16.20	16.58	34.60	33.95
ALC [26]	2.40	6.69	3.56	0.89*	0.44*	0.35	0.43	18.28	14.88	24.31	19.04
SSC [7]	1.52	4.40	2.18	-	-	9.68	17.22	2.01	5.17	33.88	33.54
TPV [23]	1.57	4.98	2.34	-	-	0.46	0.91	2.78	2.37	-	-
LRR [24]	1.33	4.98	1.59	-	-	-	-	-	-	33.67	36.01
T-Linkage [25]	0.86	5.78	1.97	-	-	-	-	-	-	-	-
S^{3}C [22]	1.94	4.92	2.61	-	-	-	-	-	-	-	-
RSIM [15]	0.78	1.77	1.01	0.68	0.70	-	-	-	-	-	-
MSSC [18]	0.54	1.84	0.83	-	-	-	0.65	**0.65**	**0.65**	-	-
Single-View											
Affine	0.40	1.26	0.59	0.15	0.10	0.25	0.35	0.93	0.82	15.76	11.52
Homography	0.45	1.61	0.71	0.18	0.10	0.70	0.48	1.23	1.08	11.45	7.14
Fundamental	1.22	7.60	1.79	1.10	0.10	5.09	2.53	4.31	3.97	13.92	5.09
Multi-View											
KerAdd	0.27	0.66	0.36	0.11	**0.00**	1.54	1.41	0.76	0.88	8.31	1.02
CoReg	0.37	0.75	0.46	**0.06**	**0.00**	0.22	**0.30**	0.83	0.73	**7.92**	0.75
Subset	**0.23**	**0.58**	**0.31**	**0.06**	**0.00**	**0.20**	**0.30**	0.77	**0.65**	8.08	**0.71**

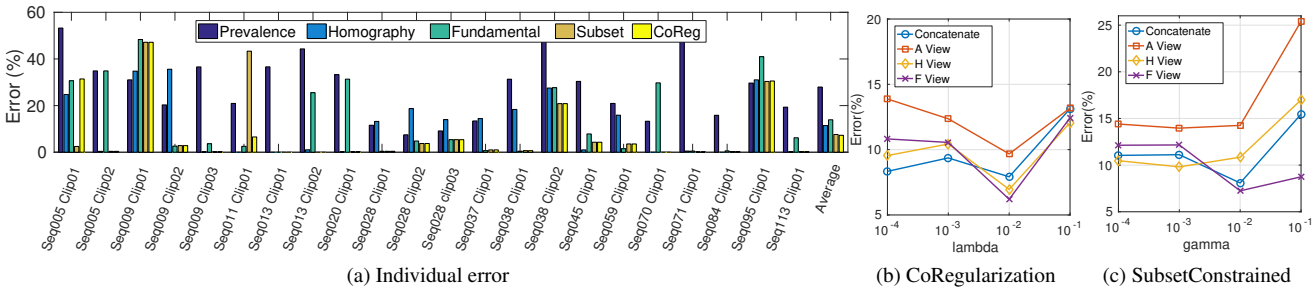
fects. While all our single-view models turned in substantially better results than the baseline approach, it is also evident from the percentage errors that each single-view model has difficulties in dealing with real-world effects. The various multi-view schemes, especially the co-regularization approach, can further improve the performance.

4.2.1 Qualitative Study

We now present the motion segmentation results on some sequences from KT3DMoSeg in Fig. 3 to better understand how different geometric models complement each other, as well as to illustrate the challenges posed by this dataset. All these sequences involve strong perspective effects in the background but the foreground moving objects often have limited depth reliefs. Many background objects have non-compact shapes, and thus the slicing effect induced by the homography/affine model is less likely to relate all the background points together due to the lower connectivity. Therefore the background tends to split in the homography view, e.g. the traffic sign in Fig. 3 (a). While fundamental matrix is more likely to discover a seamless background in theory, it is plagued by a greater susceptibility to subspace overlap in practice. For instance, the scene in Fig. 3 (b) seems to be a classic scene to which the fundamental matrix is suited, and it seems here that even though a correct fundamental matrix for the background has been estimated (manifest by the blue cluster capturing both distant points as well as the tree nearby), the overlap between the foreground cyclist and the static car means that they are wrongly grouped together. In both (a) and (b), the fusion schemes manage to correct these errors. There are also some challenges that remain in this dataset. Clearly, when the motion of the foreground object (e.g. the person in the middle of Fig. 3 (c), indicated by blue points in GroundTruth) is small or intermittent compared to that of the camera, it can be difficult to detect. Coupled with the the large depth range in the background, the algorithm can be fooled to split the background instead of segmenting the foreground. Lastly, scenes like Fig. 3 (d) still poses serious challenge. It is well known that the epipolar constraint allows a freedom to translate along the epipolar line. This allows an independent motion that is moving with respect to the background but consistent with the epipolar constraint to go undetected. In the figure, the car in front can be interpreted as a background object on the horizon, and thus the algorithm ends up splitting the big truck instead.

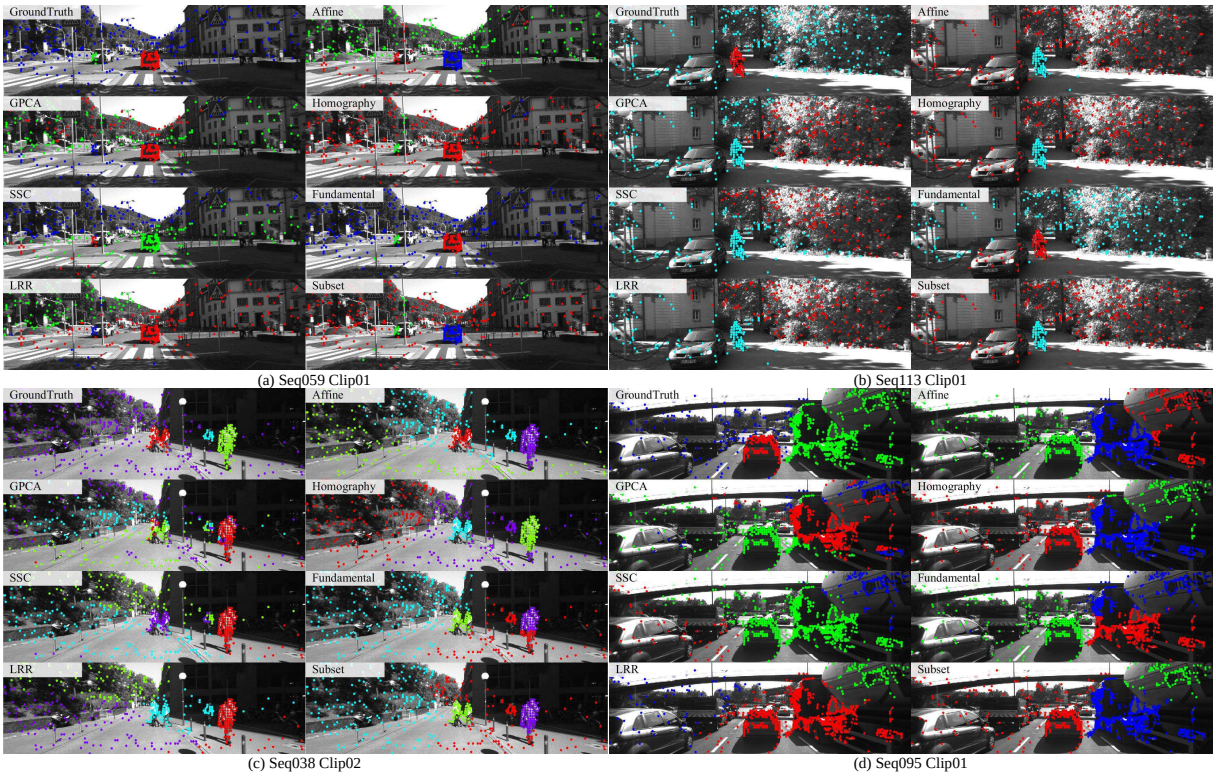
4.3. Further Analysis

Fusion Impact on Individual Views As a result of the co-regularization, each of the geometric models has their views modified; we call these the F-view, H-view, and A-view. We now analyze the performance gain experienced by these views. In particular, we investigate the performance of motion segmentation with the spectral embedding of these views after co-regularization. This is equivalent to using just a single $\mathbf{U} = \mathbf{U}_v$ for k-means clustering in the last step of Algorithm 1. The classification error over all



(a) Individual error (b) CoRegularization (c) SubsetConstrained

Figure 2: Classification error on individual sequence and sensitivity to parameters for KT3DMoSeg benchmark.



(a) Seq059 Clip01 (b) Seq113 Clip01

(c) Seq038 Clip02 (d) Seq095 Clip01

Figure 3: Examples of motion segmentation on KT3DMoSeg sequences.

KT3DMoSeg sequences v.s. λ and γ are presented in Fig. 2 (b-c). We observe from this evaluation that while the F-view (purple line) does not necessarily produce the best result compared with the H-view without co-regularization, under certain range of λ (corresponding to different coerciveness of the co-regularization), the F-view can be corrected so that its full potential is realized, producing the best of all results.

5. Conclusion

In this paper, we have contributed to an understanding of the strengths and drawbacks of homography and fundamental matrices as a geometric model for motion segmentation, not only in the extant datasets such as Hopkins155, but also for real-world sequences in KT3DMoSeg. Not only do we account for the unexpected success of the homography approach when the affinities are accumulated to over all slicing planes, we also reveal its real limitation in real-world scenes. The geometrical exactness of the fundamental matrix approach is theoretically appealing; we show how its potential can be harnessed in a multi-view spectral clustering fusion scheme. Given kernels induced from multiple types of geometric models, we evaluate several techniques to synergistically fuse them. Finally, we carry out experiments on Hopkins155, Hopkins12 and MTPV62 and achieved state-of-the-art performances on all of them. In light of the demand for real-world motion segmentation, we further propose a new dataset, the KT3DMoSeg dataset, to reflect and investigate real challenges in motion segmentation in the wild.

References

[1] T. E. Boult and L. G. Brown. Factorization-based segmentation of motions. In *Visual Motion, 1991., Proceedings of the IEEE Workshop on*, 1991. 3

[2] P. Chen. Optimization algorithms on subspaces: Revisiting missing data problem in low-rank matrix. *International Journal of Computer Vision*, 2008. 6

[3] T. Chin, H. Wang, and D. Suter. The ordered residual kernel for robust motion subspace clustering. In *NIPS*, 2009. 3, 4

[4] T.-J. Chin, J. Yu, and D. Suter. Accelerated hypothesis generation for multi-structure robust fitting. In *ECCV*, 2010. 3

[5] J. P. Costeira and T. Kanade. A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, 1998. 3

[6] R. Dragon, B. Rosenhahn, and J. Ostermann. Multi-scale clustering of frame-to-frame correspondences for motion segmentation. In *ECCV*, 2012. 1

[7] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013. 3, 6, 7

[8] C. W. Gear. Multibody grouping from motion images. *International Journal of Computer Vision*, 1998. 3

[9] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research*, 2013. 1, 6

[10] A. Goh and R. Vidal. Segmenting motions of different types by unsupervised manifold clustering. In *CVPR*, 2007. 3

[11] L. Goshen and I. Shimshoni. Guided sampling via weak motion models and outlier sample generation for epipolar geometry estimation. *International Journal of Computer Vision*, 2008. 1

[12] A. Gruber and Y. Weiss. Multibody factorization with uncertainty and missing data using the em algorithm. In *CVPR*, 2004. 3

[13] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4, 5

[14] H. C. Huang, Y. Y. Chuang, and C. S. Chen. Affinity aggregation for spectral clustering. In *CVPR*, 2012. 3

[15] P. Ji, M. Salzmann, and H. Li. Shape interaction matrix revisited and robustified: Efficient subspace clustering with corrupted and incomplete data. In *ICCV*, 2016. 6, 7

[16] H. Jung, J. Ju, and J. Kim. Rigid Motion Segmentation Using Randomized Voting. In *CVPR*, 2014. 1

[17] A. Kumar, P. Rai, and H. Daume. Co-regularized Multi-view Spectral Clustering. In *NIPS*, 2011. 3, 4, 6

[18] T. Lai, H. Wang, Y. Yan, T. J. Chin, and W. L. Zhao. Motion Segmentation Via a Sparsity Constraint. *IEEE Transactions on Intelligent Transportation Systems*, 2017. 1, 2, 3, 4, 6, 7

[19] T. Lange and J. M. Buhmann. Fusion of similarity data in clustering. In *NIPS*, 2006. 3

[20] F. Lauer and C. Schnörr. Spectral clustering of linear subspaces for motion segmentation. In *ICCV*, 2009. 3

[21] N. Lazic, I. Givoni, B. Frey, and P. Aarabi. Floss: Facility location for subspace segmentation. In *ICCV*, 2009. 3

[22] C. G. Li and R. Vidal. Structured Sparse Subspace Clustering: A unified optimization framework. In *CVPR*, 2015. 6, 7

[23] Z. Li, J. Guo, L. F. Cheong, and S. Z. Zhou. Perspective motion segmentation via collaborative clustering. In *ICCV*, 2013. 1, 6, 7

[24] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013. 3, 7

[25] L. Magri and A. Fusiello. T-linkage: A continuous relaxation of J-linkage for multi-model fitting. In *CVPR*, 2014. 6, 7

[26] S. Rao, R. Tron, R. Vidal, and Y. Ma. Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010. 3, 6, 7

[27] Y. Sugaya and K. Kanatani. Geometric structure of degeneracy for multi-body motion segmentation. In *In Workshop on Statistical Methods in Video Processing*, 2004. 1, 3

[28] N. Sundaram, T. Brox, and K. Keutzer. Dense point trajectories by GPU-accelerated large displacement optical flow. In *ECCV*, 2010. 6

[29] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 1992. 3

[30] P. Torr, A. Zisserman, and S. Maybank. Robust Detection of Degenerate Configurations while Estimating the Fundamental Matrix. *Computer Vision and Image Understanding*, 1998. 1, 3

[31] R. Tron and R. Vidal. A Benchmark for the Comparison of 3-D Motion Segmentation Algorithms. In *CVPR*, 2007. 1, 3, 6, 7

[32] R. Vidal and R. Hartley. Motion segmentation with missing data using powerfactorization and gpca. In *CVPR*, 2004. 3

[33] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (gpca). *IEEE transactions on pattern analysis and machine intelligence*, 2005. 3

[34] R. Vidal, R. Tron, and R. Hartley. Multiframe motion segmentation with missing data using PowerFactorization and GPCA. *International Journal of Computer Vision*, 2008. 3, 6, 7

[35] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 2007. 3

[36] H. Wang, F. Nie, and H. Huang. Multi-view clustering and feature learning via structured sparsity. In *ICML*, 2013. 3

[37] X. Wang, B. Qian, and I. Davidson. On constrained spectral clustering and its applications. *Data Mining and Knowledge Discovery*, 2014. 5

[38] J. Yan and M. Pollefeys. A General Framework for Motion Segmentation : Degenerate and Non-degenerate. In *ECCV*, 2006. 6, 7

[39] J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *ECCV*, 2006. 3