Fitting multiple projective models using clustering-based Markov chain Monte Carlo inference

George Terzakis¹, Phil F. Culverhouse¹, Guido Bugmann¹ and Robert Sutton²

¹Center for Robotics and Neural Systems, Plymouth University, Drake Circus PL4 8AA, Plymouth, UK

²School of Marine Science and Engineering, Plymouth University, Drake Circus PL4 8AA, Plymouth, UK

{Georgios.Terzakis, P.Culverhouse, G.Bugmann, R.Sutton}@plymouth.ac.uk

Abstract

An algorithm for fitting multiple models that characterize the projective relationships between point-matches in pairs of (or single) images is proposed herein. Specifically, the problem of estimating multiple algebraic varieties that relate the projections of 3 dimensional (3D) points in one or more views is predominantly turned into a problem of inference over a Markov random field (MRF) using labels that include outliers and a set of candidate models estimated from subsets of the point matches. Thus, not only the MRF can trivially incorporate the errors of fit in singleton factors, but the sheer benefit of this approach is the ability to consider the interactions between data points.

The proposed method (CSAMMFIT) refines the outlier posterior over the course of consecutive inference sweeps, until the process settles at a local minimum. The inference “engine” employed is a Markov Chain Monte Carlo (MCMC) method which samples new labels from clusters of data points. The advantage of this technique pertains to the fact that cluster formation can be manipulated to favour common label assignments between points related to each other by image based criteria. Moreover, although CSAMMFIT uses a Potts-like pairwise factor, the inference algorithm allows for arbitrary prior formulations, thereby accommodating the needs for more elaborate feature based constraints.

Keywords: Multiple model fitting, clustering, Markov chain Monte Carlo, two-view geometry, Markov random field.

1. Introduction

The detection of multiple projective varieties has significant practical ramifications in the fields of computer vision, robotics, pattern recognition and architecture. A significant category of such varieties are lines, conics and projectivities (homography tensors) [1, 2] which can lead to the discovery of planar surfaces and objects within scenes [3-5]. In terms of motion, multifocal tensors [6, 7] can be used to detect multiple moving objects, camera movement and degenerate configurations [8, 9]. The detection of such models in one or more views becomes an arduous task afflicted not only by the quality (or the lack thereof) of the point matches but also by the lack of prior knowledge on the nature of the outliers which may be attributed to a number of reasons (occlusions, multiple motions, changes in lighting, etc.). Several methods involving least squares fitting or iterative optimization of cost functions have been proposed for the estimation

of the parameters of these varieties. For a thorough treatment on these methods, the reader is deferred to Hartley and Zisserman [8].

Parameter estimation yields an optimal solution in terms of some distance measure with respect to all data points. This however implies that all data points indeed belong to the chosen model, an assumption which is usually not true. Thus, the task of model fitting typically includes inlier detection, given that a certain type of model exists in the data. By far, the most popular tool to achieve this, is random sampling consensus (RANSAC) [10]. The algorithm has been widely employed, not only because of its simplicity, but also because of its remarkably accurate results in a wide variety of model fitting applications.

One of the early works that attempt to capture the best model amongst the set of point matches is Phillip Torr’s MLESAC [11], a RANSAC variant that attempts to find the single optimal subset of matching point-pairs which minimize the average squared error over the entire set of matches. In the same spirit, but with a rather different approach to candidate model computation, Ondrej Chum proposed DEGENSAC [12], yet another RANSAC variant trying to find a best-fit homography for a scene that contains a dominant plane. The algorithm uses RANSAC to compute the homography of the dominant plane with respect to the two views and thereby estimates the fundamental matrix using the plane induced (virtual) parallax trick [13].

1.1 Methods for multiple model detection

While both MLESAC and DEGENSAC silently assume the existence of some unique optimal model fitting the correspondences, a method proposed by Tong [9] is not constrained by any such assumption, as it attempts to discover multiple fundamental matrices for an arbitrary set of correspondences by propagating information through tensor fields in 4D. To compute the models however, RANSAC is employed in a cascaded manner: Each model is calculated by using the set of outliers of the previous RANSAC as input to a new RANSAC execution. Cascaded RANSAC execution (also referred to as “sequential RANSAC” in literature) has been widely employed in order to cope with the existence of multiple models [9, 12, 14].

The MultiRANSAC algorithm, an actual generalization of RANSAC for multiple models, was introduced by Zulliani et al. [15]. As the name implies, the algorithm generalizes the RANSAC core concept to fitting a predetermined number of models to the data. MultiRANSAC expands the minimal sampling sets (MSS) in a cascaded manner, but the consensus sets (CS) are computed based not only on the expanded MSSs but also in terms of the CSs estimated in the previous sets. The authors present results therein superior to the ones obtained with cascaded RANSAC executions.

An algorithm that builds on the MultiRANSAC concept by introducing a clustering procedure amongst the data points following the initial determination of consensus sets, is J-linkage by Roberto Toldo and Andrea Fusiello [16]. J-linkage determines the initial consensus sets without excluding overlaps. In other words, data points may belong to more than one

consensus sets. Multiple participants (data points) are joined to form the initial clusters and the process thereafter evolves by unifying clusters with the smallest Jackard distance, thereby yielding new model parameters from their respective supports (i.e., the sets of points used to estimate the model parameters). Unlike MultiRANSAC, J-linkage can automatically conclude with the actual number of models throughout successive clustering-and-merging steps.

1.2 Inference based model detection, related work and contributions

Recently, Isack and Boykov presented PeARL, an algorithm that minimizes an energy function comprising the errors of fit and simple pairwise interactions for multiple geometric models [17]. To the best of our knowledge, this is the closest analogue to the method introduced in this paper.

In PeARL, Isack and Boykov argue based on their results, that greedy approaches to model updates via the estimation of their respective consensus sets used in J-linkage clustering, RANSAC, MultiRANSAC, or even Hough transform [18] with mean-shift [19] leave plenty of margin for model misclassifications, especially for increased levels of noise in the data. Instead of the rather greedy approaches for refinement and/or unification of consensus sets, they propose the minimization of the following energy function:

$$E(X) = \sum_r e_r(X_r) + T \sum_{s \in \mathcal{N}(r)} w_{r,s} \mathcal{I}\{X_r \neq X_s\} \quad (1)$$

where e_r is some error function regarding data point r and its respective assigned model label X_r , T is a positive constant, $\mathcal{I}$ is the indicator function, $\mathcal{N}(r)$ is the Markov blanket (neighbourhood) of r , $w_{r,s}$ is a distance-related constant such that, $w_{r,s} = \exp(-\|r - s\|^2 / \sigma^2)$ and $\|r - s\|$ is the distance between r and s . The algorithm samples random groups of points and estimates the initial parameters of a number of candidate models. It then performs repeated energy minimization sweeps by re-estimating model parameters from the minimization results. The energy function is minimized using graph-cuts [20], a method for efficient approximation of local minima. The results reported by the authors are clearly in favour of the method as opposed to J-linkage, MultiRANSAC and mean-shift in Hough space.

The energy minimization in PeARL implies probabilistic inference over a Markov network in which pairwise interactions are obtained by imposing a graph structure to the data points. In such a framework, one is able to incorporate several factors that characterize not only the distribution of the error of fit, but also the interactions between the point matches, such as distance, local gradient histogram similarities, etc., in order to make inference more “knowledgeable” of other sorts of aspects of visual content. It is worth noting here that a key limitation of graph-cuts is that they can be applied only to specific types of pairwise terms (specifically, ones that are metrics or semi-metrics). This clearly excludes a wide variety of priors which are likely to significantly improve optimization.

In this paper, we generalize the notion of energy minimization by restating the problem in terms of obtaining the joint assignment that maximizes the following probability:

$$P(X|\mathcal{M}) \propto \prod_r \varphi_r(X_r) \prod_{s \in \mathcal{N}(r)} \psi_{r,s}(X_r, X_s) \quad (2)$$

where $\mathcal{M}$ is a set of candidate (or proposed) models, X_r is the label of data point r , φ_r is a singleton factor associated with r and $\psi_{r,s}$ a pairwise factor associated with r and s . The label X_r assumes values in $\mathcal{M} \cup \{o\}$ where o denotes an outlier. Please note here that, unlike PeARL, in our formulation outliers are a valid label just like any candidate model. The existence of outliers in the label set can generally prevent the inference algorithm from overestimating “bad” models. As will be shown in the following sections, the singleton factor φ_r is modeled in such a way as to approximate the posterior of X_r given the set of ground-truth models (i.e., the most suitable set of models for the data). When this approximation is not realistic, outliers should appear in greater numbers. Finally, we introduce a slightly more elaborate prior than the indicator function which penalizes non-uniformity of labels by considering the distance between points in the context of all edge lengths in the MRF (whereas PeARL weighs the prior with a function of absolute distance).

Since the proposed distribution can be fairly generic especially in terms of choice of prior, we propose the generalized Swendsen-Wang MHMCMC algorithm (henceforth, GSW) by Barbu and Zhu [21] as the preferred inference “engine”. The results provided in their paper suggest that GSW performs generally better than graph-cuts. Most importantly, GSW is a clustering algorithm at its core and inference progresses by reorganizing the connectivity on the edges of an underlying graph structure $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of vertices corresponding to data points and $\mathcal{E}$ is a set of edges. An activation probability p_η is assigned to each edge $\eta \in \mathcal{E}$. Thus, the graph not only provides a neighbourhood system for the MRF, but also a framework of topological connections for the GSW algorithm. Formally, our sampling strategy is equivalent to sampling from the joint distribution of (2), augmented with a set of Boolean variables, $U = \{u_\eta: u_\eta = \mathcal{I}(\eta \text{ is "ON"}), \eta \in \mathcal{E}\}$, corresponding to the activation status of the edges (“ON” or “OFF”) and imposing a hard constraint for same label assignment to all variables connected with an edge that is switched “ON”:

$$P(X, U|\mathcal{M}) \propto \left(\prod_r \varphi_r(X_r) \prod_{s \in \mathcal{N}(r)} \psi_{r,s}(X_r, X_s) \right) \left(\prod_{\eta = \langle i, j \rangle \in \mathcal{E}} (1 + u_\eta(\mathcal{I}\{X_i = X_j\} - 1)) p_\eta^{\mathcal{I}\{X_i = X_j\}} (1 - p_\eta)^{1 - \mathcal{I}\{X_i = X_j\}} \right) \quad (3)$$

where $\langle i, j \rangle$ denotes the edge between the i^{th} and j^{th} vertices (data points). It can be easily seen that the marginal over U yields back the joint distribution of (2). Note here that the separate product indexes on the right in (3) allow for different structures between the MRF and the graph.

2. Method

Fig. 1. a) A set of data points and the respective Delaunay edges (Oxford buildings dataset [22]). b) A factor sub-graph of the resulting MRF for a triangle defined by vertices (data points) r , s and t .

As a first step, a graph structure $G = (\mathcal{V}, \mathcal{E})$ is obtained over the data points (Figure 1a) using Delaunay triangulation [23]; this type of triangulation ensures connectivity between closest neighbours.

To initialize iteration, a set of candidate models, $\mathcal{M}$, is estimated from respective minimal sample sets. From the candidate models, for each vertex (data point) r , an error vector $e_r = [e_{r,1} \ \dots \ e_{r,|\mathcal{M}|}]^T$ is computed. Using e_r , singleton factors ϕ_r are defined in the set of labels $\mathcal{M} \cup \{o\}$. Pairwise factors $\psi_{r,s}$ are merely penalty functions of distance between points (in the context of the graph), albeit more elaborate formulations can be used

As part of the GSW algorithm preparation, activation probabilities p_η should be specified for every edge $\eta \in \mathcal{E}$. These probabilities reflect proximity in the spatial context of the graph and therefore do not depend on the label set.

The proposed method executes inference iteratively until a global energy measure (section 3.3) has reached a local minimum. Following each inference sweep, the set of candidate models is estimated based on the current label assignments. The reader is referred to the paper by Barbu and Zhu [21] for more details on GSW cuts.

2.1 Singleton factors

As mentioned earlier, singleton factors are designed to “overestimate” the posterior of a data point r being an outlier given a set of models $\mathcal{M}$, as if this set was an “ideal” selection. In particular, we exaggerate the magnitude of the greatest likelihood of error given some model and thereby obtain an un-normalized probability measure which typically overestimates outliers for a bad set of models, but will get close to the actual posterior when these models are close to ground-truth.

The aforementioned design of singleton factors relies on modelling the probability of the error vector e_r associated with data point r , as a Gaussian mixture model contaminated with uniform noise (notion is loosely adopted by an early publication by Torr and Zisserman [11]):

$$p(e_r) = p(e_{r,1}, \dots, e_{r,|\mathcal{M}|}) = \lambda \sum_{k=1}^{|\mathcal{M}|} p(e_r|m_k)p(m_k) + (1 - \lambda)c \quad (4)$$

where $p(e_r|m_k) \sim N(0, \Sigma_k)$ and $p(m_k)$ are the error likelihood and prior of model m_k , $c > 0$ is a constant related to outliers and λ is the probability of any data point being an inlier. With a series of manipulations, the posterior probability of data point r being an outlier can be written in the following form (see appendix A for derivation):

$$P(o|e_r) = \frac{\alpha p(e_r|m_\xi)}{\sum_{n=1}^{|\mathcal{M}|} p(e_r|m_n) + \beta} \quad (5)$$

where $\xi = \operatorname{argmax}_k \{p(e_r|m_k)\}$ is the index that maximizes the error likelihood of a model label, α is related to the quality (in terms of errors of fit) of the set of candidate models and β is related to the percentage of outliers in the data points. In general, $1 \leq \alpha \leq |\mathcal{M}|$ and $\beta \geq 0$.

To define the singleton φ_r , we treat the quantities α and β as parameters. Thus, φ_r becomes the un-normalized measure of a distribution that approximates the true posterior of all labels in $\mathcal{M} \cup \{o\}$ given the error vector e_r with respect to the parameters α and β (the distribution implied by φ_r will henceforth be referred to as the α - β approximation):

$$\varphi_r(X_r) = \begin{cases} p(e_r|X_r) & , \quad X_r \in \mathcal{M} \\ \sum_{\substack{n=1 \\ n \neq \xi}}^{|\mathcal{M}|} p(e_r|m_n) + (1 - \alpha)p(e_r|m_\xi) + \beta, & X_r = o \end{cases} \quad (6)$$

Values of α in the vicinity of 1 will typically overestimate the quality of the given candidate models and, most likely, will cause many outliers to appear in the joint (i.e., the result of the GSW inference). However, as the set of candidate models is being re-computed after successive inference sweeps, the α - β approximation will move closer to the actual posterior and the joint produced by the GSW inference will become more representative of the ground truth. On the other hand, values of β close to 0 indicate a very low percentage of outliers in the data points, an assumption which typically is true (but not always).

The error likelihood $p(e_r|m_k)$ is a multivariate Gaussian, but it depends only on the k -th error component, $e_{r,k}$. Formally, the information matrix of the likelihood has zero elements everywhere except for the k -th diagonal entry:

$$\Omega_k = \sigma_k^{-2} \operatorname{diag}\{a_i\},$$

where $a_i = (1 + (-1)^{\mathcal{I}\{i-k\}})/2$, $i = 1, \dots, |\mathcal{M}|$ and σ_k^2 is the variance. Consequently, the likelihood is a degenerate multivariate (in practice, univariate) normal distribution:

$$p(e_r|m_k) = \left(\frac{1}{2\pi} \det^\dagger \Omega_k\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2} e_r^T \Omega_k e_r\right) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2} \frac{(e_{r,k})^2}{\sigma_k^2}\right) \quad (7)$$

where $\det^\dagger A$ denotes the pseudo-determinant of A.

2.2 Pairwise factors

Pairwise factors impose label uniformity considering distances between vertices in the context of the graph. Specifically, a pairwise factor $\psi_{r,s}$ is given by the following expression:

$$\psi_{r,s}(X_r, X_s) = \begin{cases} \exp(-T_1 p_{<r,s>}) & , \quad X_r = X_s \\ \exp(T_2 p_{<r,s>}) & , \quad X_r \neq X_s \end{cases} \quad (8)$$

where $p_{<r,s>}$ is an activation probability for the edge $<r, s>$ (see section 2.3) and T_1, T_2 are positive “temperature” constants (in practice, $T_1 \ll T_2$ or $T_1 = 0$). Since $p_{<r,s>}$ reflects distance in relation to other edges in the graph, the penalty is context aware. Clearly, considering the rich informational content of images, suggestions for a more elaborate prior can be limited only by imagination.

2.3 Activation probabilities for the graph edges

The GSW algorithm produces samples from the joint of (3). It is therefore necessary, except new labels, to produce new edge activation states at each step of the sampling process. To do so, the algorithm activates or deactivates edges that have same label assignment on their vertices by some probability p_η or $1 - p_\eta$ respectively. Activation probabilities do not depend on label assignments, but can be related to the edges and/or the image local/global features. Since new samples are assigned to groups of variables that belong to graph clusters, it would be reasonable for the activation probabilities to reflect distance or other image-based similarity criteria so that data points that belong to the same model seek to be clustered together.

For the method introduced in this paper, edge activation probabilities are designed to reflect distance in the context of the graph. In other words, we seek to establish a soft distinction between edge vertices that are “close” to each other and edge vertices that are “far” from each other, in relation to all edge distances in the graph. This can be regarded as the problem of fitting a Gaussian discriminative model over 1D data; typically, this is done with the expectation maximization algorithm for two classes, C_0 for small and C_1 for large distances. The algorithm is initialized with two Gaussians of equal variance with means set to the minimum and maximum edge length respectively (see Figure 2) so that the EM iteration “pulls” these means more towards these extremes rather than risking settlement in some local minimum closer to a median

point. Following convergence, the activation probabilities can be obtained as the joint probability of the length of an edge and class C_0 , $p(\|r - s\|, C_0)$:

$$p_{<r,s>} = p(\|r - s\|, C_0) = p(\|r - s\| | C_0) \pi_0 \quad (9)$$

where $p(\|r - s\| | C_k), k \in \{0,1\}$ is the likelihood of the edge length given that the edge belongs to class C_k and π_k is the class prior. In detailed steps, the algorithm that computes the edge probabilities is as follows (EM steps are omitted):

Computation of edge activation probabilities p_η

Input: a) Set of distances $D(\mathcal{E}) = \{d_\eta: d_\eta = \|r - s\|, \eta = \langle r, s \rangle \in \mathcal{E}\}$,

Output: $p_{<r,s>} = p(\|r - s\|, C_0)$

1. Initialize the EM algorithm with the following parameters: Two classes C_0 and C_1 with respective means $\mu_0 = \min D(\mathcal{E})$ and $\mu_1 = \max D(\mathcal{E})$ and variances $\sigma_1^2 = \sigma_2^2 = (\mu_1 - \mu_0)/2$; the class priors π_0 and π_1 are set to $1/2$. The marginal probability of length is therefore given by,

$$p(\|r - s\|) = \underbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{(\|r - s\| - \mu_0)^2}{2\sigma_0^2}\right)}_{p(\|r-s\| | C_0)} \pi_0 + \underbrace{\frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(\|r - s\| - \mu_1)^2}{2\sigma_1^2}\right)}_{p(\|r-s\| | C_1)} \pi_1$$

2. Run EM until some convergence criterion is met.
3. Obtain the edge probability $p_{<r,s>} = p(\|r - s\|, C_0)$.

Fig. 2. a) Delaunay edges on the second view of Merton1 (Oxford buildings dataset [22]). b) The initial mixture (we choose the greatest length as initial variance, $\sigma_0^2 = \sigma_1^2 = 459.24$ for numerical stability and the means $\mu_0 = 3.5726$ and $\mu_1 = 459.2447$ as the smallest and greatest edge length) and the mixture after EM training ($\mu_0 = 36.5307$, $\mu_1 = 128.6370$ and variances $\sigma_0^2 = 404.8208$, $\sigma_1^2 = 6635.5$).

2.4 Proposal distribution and acceptance ratio for the GSW algorithm

Our choice of proposal distribution for the GSW algorithm results in a sampling process which can be regarded as a Gibbs sampler using $p(U|X)$ and $p(X|U)$ from (3). In the context of the GSW algorithm this is equivalent to the following acceptance probability:

$$a(S_t \rightarrow S_{prop}) = \min \left\{ \frac{\prod_{\eta \in \mathcal{E}(Y \leftrightarrow (\mathcal{V}_{prop}(m_{prop}) - Y))} (1 - p_\eta) q(X_Y = m_Y | S_{prop}, \mathcal{M}) P(X_{prop} | \mathcal{M})}{\prod_{\eta \in \mathcal{E}(Y \leftrightarrow (\mathcal{V}_t(m_Y) - Y))} (1 - p_\eta) q(X_Y = m_{prop} | S_t, \mathcal{M}) P(X_t | \mathcal{M})}, 1 \right\} \quad (10)$$

where $S_t = (X_t, U_t)$ is the state of the graph at time t , $\mathcal{E}(A \leftrightarrow B)$ is the set of edges between the sets of vertices A and B, $\mathcal{V}_t(m)$ is the subset of vertices to which the label m is assigned, Y is a selected connected component in the graph, $P(X|\mathcal{M})$ is the target distribution as given in (2) and,

$$q(X_Y = m | S_t, \mathcal{M}) = P(X_Y = m, X_{\mathcal{V}_t - Y} | \mathcal{M}) \quad (11)$$

where $X_{\mathcal{V}_t - Y}$ denotes the joint assignment to all data points except the ones belonging to Y . Obviously, $X_{\mathcal{V}_t - Y} = X_{\mathcal{V}_{prop} - Y}$ as a direct consequence of the fact that we only change the labels of data points in Y . Hence, the two right-most fractions of the ratio in (10) will cancel out, yielding the following simplified expression for the acceptance probability:

$$a(S_t \rightarrow S_{prop}) = \min \left\{ \frac{\prod_{\eta \in \mathcal{E}(Y \leftrightarrow (\mathcal{V}_{prop}(m_{prop}) - Y))} (1 - p_\eta)}{\prod_{\eta \in \mathcal{E}(Y \leftrightarrow (\mathcal{V}_t(m_Y) - Y))} (1 - p_\eta)}, 1 \right\} \quad (12)$$

3. Clustering based SAMpling for Multiple Model FITting (CSAMMFIT): Intuition and tuning

Given a set of data points $\mathcal{V}$, the set of all possible models Ω is a space of combinations with size complexity $O(|\Omega|) = 2^{|\mathcal{V}|}$. CSAMMFIT searches Ω by starting from a random subset and throughout successive inference sweeps, it discards and/or recombines models until it settles at a local minimum with respect to a global energy measure.

Let $\mathcal{M}_0 \subseteq \Omega$ be the initial set of candidate models. There are many ways to obtain $\mathcal{M}_0$; typical strategies involve the classic RANSAC [21] approach that involves the sampling of random points in $\mathcal{V}$, or slightly more elaborate sampling schemes such as the one used in J-linkage [16]. Generally, at iteration- t , given a set of candidate models $\mathcal{M}_t$, the algorithm constructs the respective factors (see sections 2.1 and 2.2) and executes a GSW inference sweep. Following a sufficient number of Markov chain transitions, the assigned labels define the new supports of the models in $\mathcal{M}_t$. The algorithm discards the models with no (or very little support),

a step which can be thought of as regularization in terms of number of models and thereafter re-estimates the parameters of each remaining model from their new supports. Finally, if some of the re-estimated model parameters differ from others by a distance less than a threshold, the respective supports are fused and they are replaced by a single model in $\mathcal{M}_{t+1}$. The process repeats until the change in a global energy measure is close to zero, or a maximum number of iterations has been reached.

From the above, at a fairly abstract level, the algorithm can be described as follows:

Clustering based Sampling for Multiple Model FITting (CSAMMFIT)

Input: a) Set of data points $\mathcal{V}$, b) Model parameter vector (model type), c) Number of initial models, $|\mathcal{M}_0|$, d) Maximum number of iterations, $maxIter$.

Output: Optimal set of models $\mathcal{M}_{opt} \subseteq \Omega$ and respective joint assignment $X_{opt} \sim P(X|\mathcal{M}_{opt})$.

1. Compute $\mathcal{M}_0$ by sampling random points in $\mathcal{V}$.
2. Generate a graph $G = (\mathcal{V}, \mathcal{E})$ and compute an edge activation probability p_η for each $\eta \in \mathcal{E}$.
3. $t \leftarrow 0$. $E_t \leftarrow 0$. $E_{t-1} \leftarrow 0$. $\Delta E_t \leftarrow \infty$.
4. Repeat until $t \geq maxIter$ or $\Delta E_t \approx 0$:
5. Using the edges of G as a neighborhood system and the label set $\mathcal{M}_t \cup \{o\}$ compute singleton and pairwise factors as shown in equations (6) and (8).
6. Run the generalized Swendsen-Wang algorithm for the joint probability distribution defined by the singleton and pairwise factors, using edge activation probabilities p_η in order to sample new graph partitions.
7. $E_{t-1} \leftarrow E_t$.
8. Compute E_t as a measure proportional to the un-normalized negative log likelihood.
9. $\Delta E_t = E_t - E_{t-1}$.
10. Create a set of models $\overline{\mathcal{M}}_t$ by selecting the models in $\mathcal{M}_t$ which have non-empty support and re-estimate their parameters.
11. For each model $m \in \overline{\mathcal{M}}_t$, find the subset $M(m, \varepsilon) \subseteq \overline{\mathcal{M}}_t$ such that for every model $l \in M(m, \varepsilon)$ the parameter vector of m differs from the parameter vector of l no more than ε . In other words, if we denote the parameter vector of m by $\theta(m)$, then $\|\theta(m) - \theta(l)\| < \varepsilon$ by a given metric.
12. $\mathcal{M}_{t+1} = \emptyset$.
13. For each class in partition $\{M(m_1, \varepsilon), M(m_2, \varepsilon), \dots, M(m_{|\mathcal{M}|}, \varepsilon)\}$, generate a single model by unifying the supports of the models that belong to the class and adding the resulting set to $\mathcal{M}_{t+1}$.
14. $t \leftarrow t + 1$.

3.1 Selecting the initial set of candidate models

The initial set of candidate models can be sampled in many different ways, given the nature of the models. For instance, it seems prudent to sample two random points from the dataset in order to obtain a line, yet if we sample four points uniformly at random, the resultant homography will have a very poor impact on the likelihood. On the other hand, if one chooses to obtain the minimal sampling set of a homography by randomly choosing a center point and then adding its

three closest neighbors, the resulting model will have a greater likelihood than most homographies obtained by fitting the parameters to four random data points. In general, sampling strategies may vary depending on the type of model.

3.2 Obtaining a new set of models

In theory, the GSW inference maximizes the joint X . Using the label assignments $X_r \in \mathcal{M}_t$, the supporting sets $\mathcal{S}(m) \subseteq \mathcal{V}$ for every $m \in \mathcal{M}$ are obtained. From the supports, a new set of models is obtained as,

$$\overline{\mathcal{M}}_t = \left\{ \overline{m} \in \Omega: \theta(\overline{m}) = \underset{\theta}{\operatorname{argmin}} f(\theta, \mathcal{S}(m)), m \in \mathcal{M}_t \right\} \quad (13)$$

where $f(\theta, S)$ is an error function expressing goodness of fit of the parameter vector θ to the point set S , Ω is the space of all possible models that can be obtained from $\mathcal{V}$ and $\theta(m)$ is the parameter vector associated with model m . In other words, $\overline{\mathcal{M}}_t$ is the set of new models obtained by re-estimating the models in $\mathcal{M}_t$ that end up having non-empty support following the inference sweep. It is evident that the re-estimation of the original models can only improve a global likelihood measure, as pointed out in [17].

Another factor that must be accounted for during the re-estimation of candidate models, is the multiplicity of entries in $\overline{\mathcal{M}}_t$ that correspond to the same model. We dub these entries *spurious* models. To account for spurious models, a distance metric $\|\cdot\|$ in parameter space is considered. Such metrics can be geodesics on the manifolds on which the model parameters lie. Given a threshold ε , one can define a simple equivalence relationship using the following classes for each $\overline{m} \in \overline{\mathcal{M}}_t$:

$$M(\overline{m}, \varepsilon) = \{l \in \overline{\mathcal{M}}_t: \|\theta(\overline{m}) - \theta(l)\| < \varepsilon\} \quad (14)$$

Having obtained the set of classes, $C = \{M(\overline{m}_1, \varepsilon), M(\overline{m}_2, \varepsilon), \dots, M(\overline{m}_{|\mathcal{M}|}, \varepsilon)\}$, which is a partition of $\overline{\mathcal{M}}_t$, the new set of candidate models $\mathcal{M}_{t+1}$ is obtained by computing the parameters of a single model for each equivalence class, using the combined supports of the models that belong to the class:

$$\mathcal{M}_{t+1} = \left\{ l_i \in \Omega: i \in \{1, \dots, |C|\}, \theta(l_i) = \underset{\theta}{\operatorname{argmin}} f\left(\theta, \bigcup_{\overline{m} \in M(\overline{m}_i, \varepsilon)} \mathcal{S}(\overline{m})\right) \right\} \quad (15)$$

3.3 A global energy measure

Although the un-normalized negative log-likelihood (energy) of each GSW sweep is a valid quality measure of convergence, it cannot however be used for comparison between successive inferences. This is because each inference sweep concerns a distinct set of factors, corresponding to a distinct set of candidate models. To obtain a measure of convergence which is independent

of the formulations of the GSW sweeps, one is required to consider a more global view of the problem.

Formally stated, the distribution of label assignments to a data point r in terms of the sample space Ω and its subsets can be expressed as the following marginal over all possible model subsets:

$$P(X_r) = \begin{cases} \sum_{\mathcal{M} \in \mathcal{P}(\Omega)} P(X_r, \mathcal{M}), & X_r \neq o \\ 1 - \sum_{\mathcal{M} \in \mathcal{P}(\Omega)} P(X_r, \mathcal{M}), & X_r = o \end{cases} \quad (16)$$

where $\mathcal{P}(\Omega)$ is the power set of Ω .

In this global model, it is reasonable to assume independence between any two labels X_r and X_q , since all possible model subsets are taken into consideration in the marginal distribution and therefore, the singletons should be “strong enough” to beat the effects of any interactions in the joint:

$$P(X_r) = \prod_{r \in \mathcal{V}} \sum_{\mathcal{M} \in \mathcal{P}(\Omega) \cup \{o\}} P(X_r, \mathcal{M}) \quad (17)$$

Following several manipulations on (17) (see Appendix B for derivations and rationale), we conclude that maximizing $P(X_r)$ depends mainly on the maximization of the joint $P(X_r, \mathcal{M})$. The latter can be expressed in the following regularized likelihood measure:

$$E(X, \mathcal{M}) = - \left[\sum_{r \in \mathcal{V}} \log P(X_r | \mathcal{M}) - |\mathcal{V}| \sum_{m \in \mathcal{M}} \log \left(\frac{|\mathcal{V}|}{|\mathcal{S}(m)|} \right) \right] \quad (18)$$

A quick examination of (18) reveals that the energy does not concern outliers as defined for use in the GSW sweeps. The likelihood term refers to labels in $\mathcal{M}$ only. To cope with this in practice, for the case that some point is labeled outlier by inference, then the worst model posterior is used in the formula. Evidently, the energy measure of (18) resembles the Bayesian information criterion (BIC) [24] in that the first summation is the negative log-likelihood, while the second summation is a regularization term that penalizes the number of models inversely proportionally to the respective number of points in the support sets.

4. Results

Since PeARL [17] is the closest known method to CSAMMFIT, we used the Oxford buildings dataset [22] as a benchmark for loose comparison with it.

Typically, neighboring points in two-view geometry support the same projectivity, whereas randomly sampled points are unlikely to verify the same model. Hence, a reasonable strategy to obtain the initial set of candidate models would be to randomly sample a point and thereafter, obtain the 3 closest neighbors in the graph.

4.1 Projectivities

Figure 3 illustrates plane detection using views 1 and 2 of the first Merton College scene (Merton 1) in the Oxford buildings dataset [22]. Figures 3a and 3b display the results obtained with CSAMMFIT; in order to loosely “emulate” PeARL without using graph cuts, we obtained results (Figure 4) using singleton factors with a very large value for a (i.e., outlier labels have very low probability) and the Potts prior of (2).

a) Results with CSAMMFIT.

b) Global energy

Fig. 3. Multiple plane detection in Merton 1, views 1 and 2 (red circles denote outliers, while letters correspond to the various planes recovered in the scene). a) Results using the CSAMMFIT. b) The respective global energy per execution step in log scale.

Typical parameter values were, $\alpha = 1.1$, $\beta = 0$, $T_1 = 5 \cdot 10^3$ and $T_2 = 0$, number of initial candidate planes, 40 (4 points used for homography estimation), the threshold for model fusion was 0.01 using the Euclidean distance (Frobenius norm) between normalized homography matrices (normalization constant is obtained from the singular value decomposition of the matrix [25]), tolerance for the logarithm of the energy measure was 0.004 and maximum number of iterations, 15. The GSW algorithm is always employed for inference sweeps of 800 Markov chain transitions. The rationale behind this selection relies on the results by Barbu and Zhu[21] favoring the GSW algorithm against graph cuts.

a) Results using PeARL.

b) Global energy

Fig. 4. Multiple plane detection in Merton 1, views 1 and 2; a) Results using the energy formulation of PeARL. b) The respective global energy per execution step in log scale.

a) Results using CSAMMFIT.

b) Global energy

Fig. 5. Multiple plane detection in Merton 1, views 2 and 3; a) Results using CSAMMFIT; b) The respective global energy per execution step in log scale.

We executed the algorithm with the PeARL energy based configuration using the same parameters that we used with our method. It becomes obvious that PeARL overestimates spurious models, although the energy levels achieved are slightly worse than the ones achieved by CSAMMFIT. At the end of the process, the results obtained with PeARL contain a great deal of spurious models (the few outliers appear as a consequence of using a near-to-but-not-zero outlier posterior of 10^{-18}). This suggests that CSAMMFIT is more robust to lower fusion thresholds than PeARL, owed to the underlying cluster based inference and the use of a context-aware prior.

Results for views 2 and 3 of Merton 1 (Figure 5 and 6) again suggest that PeARL's energy formulation performs less effectively in grouping similar models, hence the squared errors dominate energy at the expense of the regularization term.

Fig. 6. Multiple plane detection in Merton 1, views 2 and 3; a) Results using the energy formulation of PeARL; b) The respective global energy per execution step in log scale.

Except for energy formulations, the effects of priors in the results were explored. Figures 7 and 8 depict the results obtained with the CSAMMFIT prior and the PeARL prior respectively for views 1 and 2 of Merton 2. As mentioned earlier, the PeARL prior considers the distance between two nodes, but only in absolute terms, whereas the proposed prior uses the edge activation probability to regulate the pairwise penalty. The temperature was increased to $T_1 = 5 \cdot 10^4$ in order to slightly amplify the effects of prior-imposed grouping in the results.

Fig. 7. Multiple plane detection in Merton 2, views 1 and 2; a) Results using the CSAMMFIT prior; b) The respective global energy per execution step in log scale.

Fig. 8. Multiple plane detection in Merton 2, views 1 and 2; a) Results using the PeARL prior; b) The respective global energy per execution step in log scale.

Please note here that the algorithms will always require initial tuning of a few parameters (namely, α , β , T and the threshold used in model fusion). We note however that CSAMMFIT yields good results for a wide variety of settings, while the PeARL energy appears to be very sensitive to temperature changes. Moreover, the PeARL prior needs to be “hand-tuned” every time a new pair of images is used; this is most likely due to the fact that coefficients $w_{r,s}$ reflect absolute distances. Model fusion is also a process that CSAMMFIT seems to be more robust than PeARL, since it generally produces uniform label assignments just throughout the inference process, while fusion occurs rarely. Figures 4 and 6 suggest that spurious models are generally resilient to inference in PeARL and therefore, in order to eliminate them, one must increase the fusion threshold. Table 1 reports success rates for the Oxford dataset without counting spurious models as misses; outliers on the other hand do count.

	CSAMMFIT		PeARL	
	Views 1 and 2	Views 2 and 3	Views 1 and 2	Views 2 and 3
Merton 1	91.12%	97.39%	88.4%	91.91%
Merton 2	99.13%	98.98%	91.28%	89.80%
Wadham	95.22%	87.99%	92.93%	93.29%

Table. 1. Results for the Oxford buildings dataset. Columns 1 and 2 correspond to the results of the proposed method for pairs of views 1, 2 and 2, 3; columns 3 and 4 correspond to the results obtained with the PeARL energy formulation (GSW inference used).

4.2 Multiple motions

Multiple motions in two views are similarly characterized by bilinear relationships between the tracked points. Thus, it is possible to formulate singletons based on some error function that evaluates model fitness, typically the Sampson distance [8]. For the sake of completeness, we report results obtained from several different pairs of views taken from two sequences in the Hopkins 155 dataset [26]. Illustrative results are shown in Figures 9 and 10.

Fig. 9. Moving car in front of a moving camera (cars1). Results obtained using CSAMMFIT on a pair of views in the sequence (frames 1, 12).

Table 2 reports success rates on two different pairs of frames in sequences *cars1* and *cars2_06* of the Hopkins 155 dataset.

	CSAMMFIT		PeARL	
	Frames 1 and 12	Frames 1 and 7	Frames 1 and 12	Frames 1 and 7
Cars1	94.38%	89.23%	91.67%	87.69%
Cars2_06	88.62%	92.68%	91.06%	90.24%

Table 2. Indicative results for the Hopkins 155 dataset. Comparison between CSAMMFIT and PeARL for pairs of frames 1, 12 and 1, 7 of sequences *cars1* and *cars2_06*.

Fig. 10. A sequence with two moving cars (cars2_06). Results obtained using CSAMMFIT on a pair of views in the sequence (frames 1, 12).

5. Conclusion

An algorithm that fits multiple projective models (or generally, algebraic varieties) to a set of data points matched across views was introduced in this paper. The algorithm treats the problem in terms of probabilistic inference over a MRF, generated by a graph structure obtained from the data points. Inference is performed using the MCMC Metropolis-Hastings generalized Swendsen-Wang clustering for arbitrary distributions. The underlying idea is to, not only consider the errors of fit, but also consider interactions between data points in the image(s), such as distance, local colour or gradient histogram differences, etc. and specify suitable priors. A global measure of likelihood (i.e., it characterizes the goodness of fit in an absolute manner) is also devised to evaluate the convergence of the process.

The method is novel to the best of our knowledge, since, with the exception of PeARL, the existing algorithms (RANSAC, MultiRANSAC, J-linkage) that attempt to resolve multiple models do not face the problem in terms of a joint posterior distribution (i.e, priors are absent). On the other hand, PeARL runs graph-cut optimization for an energy function using an error based singleton factor and a simple Potts prior for label uniformity.

The results reported by Isack and Boykov are all in favour of PeARL, so it is logical to compare CSAMFIT with their method. Please note here that detection of multiple varieties in single or multiple view geometry is a process that is inherently parametric. We therefore argue that there can be no absolutely rigorous comparison measure with other methods, since they also perform descent towards local minima, a process which may be affected by numerous parameters. For instance, the number of initial candidate models is clearly a decisive factor affecting how close to the ground truth will the process eventually get. It is possible however, one method to achieve lower energy for a number of initial models, while another that achieves slightly greater energy, requires a significantly smaller number of such models. The trade-off obviously favours the algorithm that achieves results of slightly lower quality but with fewer initial models. Other examples of such differences in process “tuning” include the numerical methods for error and model parameter estimation, the threshold and respective metric in parameter space used for model fusion and the temperature in the prior term.

An essential limitation of PeARL is that graph-cuts work only with certain types of priors (metrics or semi-metrics); hence, it cannot generalize to arbitrary types of constraints, whereas the GSW inference used by CSAMMFIT allows for arbitrary prior formulations; moreover, the proposed CSAMMFIT prior penalizes discontinuities and/or uniformity by considering distances between points in the context of the graph, while PeARL does this only in absolute terms. Also, the PeARL singletons assume labels only from the set of candidate models, while the α - β approximation (i.e., the singleton proposed in this paper) models outliers explicitly; results suggest that this formulation limits the number of spurious models, since inference avoids the weakest models and assigns outlier labels to the respective data points.

Although CSAMMFIT performs well on the Oxford dataset for various parameter settings, it is possible that entirely different circumstances and/or model types may require a

certain amount of parameter “tuning”. Parameters α and β heuristically are chosen to be 1.1 and 0 respectively, but in environments with greater numbers of outliers, different values could converge closer to the true minimum. With respect to pairwise interactions, the temperature parameter can generally vary from $5 \cdot 10^3$ to $5 \cdot 10^5$ without significant changes in the results from the Oxford dataset. However, such a variation may cause significant changes in the results when pursuing different types of models in another pair of images. Finally, model fusion accuracy greatly depends on the choice of metric over the parameter space. If the manifold is highly non-linear, Euclidean metrics can easily mislead the fusion process. CSAMMFIT showed great tolerance to spurious models, thereby significantly reducing the need for fusion. The latter suggests that the algorithm can descent “optimally” (i.e., without having to fuse spurious models) given a set of initial models of reasonable quality.

References

- [1] J. G. Semple and G. T. Kneebone, *Algebraic projective geometry*: Clarendon Press, 1952.
- [2] N. Apostoloff and A. Fitzgibbon, "Learning spatiotemporal T-junctions for occlusion detection," 2005, pp. 553-559 vol. 2.
- [3] B. M. Bell and F. W. Cathey, "The iterated Kalman filter update as a Gauss-Newton method," *Automatic Control, IEEE Transactions on*, vol. 38, pp. 294-297, 1993.
- [4] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 509-522, 2002.
- [5] M. J. Black, D. J. Fleet, and Y. Yacoob, "A framework for modeling appearance change in image sequences," 1998, pp. 660-667.
- [6] Q. T. Luong and O. D. Faugeras, "The fundamental matrix: Theory, algorithms, and stability analysis," *International Journal of Computer Vision*, vol. 17, pp. 43-75, 1996.
- [7] R. I. Hartley, "Lines and points in three views and the trifocal tensor," *International Journal of Computer Vision*, vol. 22, pp. 125-140, 1997.
- [8] R. Hartley, A. Zisserman, and I. ebrary, *Multiple view geometry in computer vision* vol. 2: Cambridge Univ Press, 2003.
- [9] W. S. Tong, C. K. Tang, and G. Medioni, "Simultaneous two-view epipolar geometry estimation and motion segmentation by 4D tensor voting," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, pp. 1167-1184, 2004.
- [10] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, pp. 381-395, 1981.
- [11] P. H. S. Torr and A. Zisserman, "MLESAC: A new robust estimator with application to estimating image geometry," *Computer vision and image understanding*, vol. 78, pp. 138-156, 2000.

- [12] O. Chum, T. Werner, and J. Matas, "Two-view geometry estimation unaffected by a dominant plane," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, pp. 772-779.
- [13] B. Boufama and R. Mohr, "Epipole and fundamental matrix estimation using virtual parallax," in *Computer Vision, 1995. Proceedings., Fifth International Conference on*, 1995, pp. 1030-1036.
- [14] H. Boulaassal, T. Landes, P. Grussenmeyer, and F. Tarsha-Kurdi, "Automatic segmentation of building facades using terrestrial laser data," *International Archives of Photogrammetry, Remote Sensing and Spatial Information Systems*, pp. 65-70, 2007.
- [15] M. Zuliani, C. S. Kenney, and B. Manjunath, "The multiransac algorithm and its application to detect planar homographies," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, 2005, pp. III-153-6.
- [16] R. Toldo and A. Fusiello, "Robust multiple structures estimation with j-linkage," *Computer Vision—ECCV 2008*, pp. 537-547, 2008.
- [17] H. Isack and Y. Boykov, "Energy-based geometric multi-model fitting," *International Journal of Computer Vision*, vol. 97, pp. 123-147, 2012.
- [18] A. K. Agrawal and R. Chellappa, "Robust ego-motion estimation and 3D model refinement using depth based parallax model," 2004, pp. 2483-2486 Vol. 4.
- [19] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: Design of dictionaries for sparse representation," *Proceedings of SPARS*, vol. 5, pp. 9-12, 2005.
- [20] H. Auvinen, T. Leppälampi, J. Taipale, and M. Teplykh, "Scale Invariant Feature Transform."
- [21] A. Barbu and S. C. Zhu, "Generalizing Swendsen-Wang to sampling arbitrary posterior probabilities," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, pp. 1239-1253, 2005.
- [22] A. Ayvaci, M. Raptis, and S. Soatto, "Occlusion detection and motion estimation with convex optimization," *Advances in Neural Information Processing Systems*, 2010.
- [23] D. T. Lee and B. J. Schachter, "Two algorithms for constructing a Delaunay triangulation," *International Journal of Parallel Programming*, vol. 9, pp. 219-242, 1980.
- [24] L. Baumela, L. Agapito, P. Bustos, and I. Reid, "Motion estimation using the differential epipolar equation," 2000, pp. 840-843 vol. 3.
- [25] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry, "An Invitation to 3D Vision. 2004," ed: Springer.
- [26] R. Tron and R. Vidal, "A benchmark for the comparison of 3-d motion segmentation algorithms," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 2007, pp. 1-8.

Appendix A: Rationale behind the α - β approximation (outlier posterior parameterization)

Using Bayes' rule the posterior probability of a model m_k given e_r is:

$$P(m_k|e_r) = \lambda \frac{p(e_r|m_k)P(m_k)}{p(e_r)} \quad (19)$$

Using (4) and (19), the posterior of an outlier is obtained:

$$P(o|e_r) = 1 - \lambda \sum_{k=1}^{|\mathcal{M}|} \frac{p(e_r|m_k)P(m_k)}{p(e_r)} = 1 - \lambda \sum_{k=1}^{|\mathcal{M}|} \frac{p(e_r|m_k)P(m_k)}{\lambda \sum_{n=1}^{|\mathcal{M}|} p(e_r|m_n)p(m_n) + (1-\lambda)c} \quad (20)$$

We now drop data point indexing for simplicity and introduce the notation $l_k = p(e|m_k)$ for the k -th error likelihood. Assuming a uniform model prior $(m_k) = \frac{1}{|\mathcal{M}|}$, the expression in (20) can be rewritten as follows:

$$\begin{aligned} P(o|e_r) &= 1 - \lambda \frac{1}{|\mathcal{M}|} \sum_{k=1}^{|\mathcal{M}|} \frac{l_k}{\lambda \frac{1}{|\mathcal{M}|} \sum_{n=1}^{|\mathcal{M}|} l_n + (1-\lambda)c} \\ \Leftrightarrow P(o|e_r) &= 1 - \sum_{k=1}^{|\mathcal{M}|} \frac{l_k}{\sum_{n=1}^{|\mathcal{M}|} l_n + |\mathcal{M}| \frac{(1-\lambda)}{\lambda} c} = 1 - \frac{1}{\sum_{n=1}^{|\mathcal{M}|} l_n + |\mathcal{M}| \frac{(1-\lambda)}{\lambda} c} \sum_{\kappa=1}^{|\mathcal{M}|} l_{\kappa} \end{aligned} \quad (21)$$

Let now $\xi = \operatorname{argmax}_k \{l_k\}$ be the index of the greatest likelihood. By factoring l_{ξ} out of the sum in the numerator in (21), the posterior assumes the following form:

$$P(o|e_r) = 1 - \frac{l_{\xi}}{\sum_{n=1}^{|\mathcal{M}|} l_n + |\mathcal{M}| \frac{(1-\lambda)}{\lambda} c} \sum_{\kappa=1}^{|\mathcal{M}|} \frac{l_{\kappa}}{l_{\xi}} \quad (22)$$

Setting $\alpha = \sum_{\kappa=1}^{|\mathcal{M}|} \frac{l_{\kappa}}{l_{\xi}}$ and $\beta = |\mathcal{M}| \frac{(1-\lambda)}{\lambda} c$, equation (22) yields:

$$P(o|e_r) = 1 - \frac{\alpha l_{\xi}}{\sum_{n=1}^{|\mathcal{M}|} l_n + \beta} \quad (23)$$

Quantity α is lower bounded by 1 and upper bounded by $|\mathcal{M}|$. It characterizes the quality of a set of candidate models; if this set is close to (or contains) the ground truth models, then, naturally, the likelihood associated to one specific model will be much greater than all other likelihood terms, hence α will be very close to 1. Quantity β can be considered as a measure of outlier presence in the population of data points.

Clearly, parameters α and β are heuristically selected and can be fine-tuned throughout consecutive inference sweeps. However, a fairly reasonable initial choice would be $\alpha = 1$ and $\beta = 1$. Choosing $\alpha = 1$ would imply that the quality of the solution is far better than the one of any other candidate, something which is generally true (but not always). On the other hand, choosing $\beta = 0$ implies that $\lambda \gg 1 - \lambda$ and therefore the probability of a point being an inlier is very high, which also is true in many sequences (but not a general rule).

Appendix B: Devising a regularized global energy measure

Formally stated, the distribution of label assignments to a data point r , in terms of the sample space Ω and its subsets can be expressed as the following marginal over all possible model subsets:

$$P(X_r) = \begin{cases} \sum_{\mathcal{M} \in \mathcal{P}(\Omega)} P(X_r, \mathcal{M}), & X_r \neq o \\ 1 - \sum_{\mathcal{M} \in \mathcal{P}(\Omega)} P(X_r, \mathcal{M}), & X_r = o \end{cases} \quad (24)$$

where $\mathcal{P}(\Omega)$ is the power set of Ω .

In this global model, it is reasonable to assume conditional independence between any two labels X_r and X_q , since all possible model subsets are taken into consideration in the marginal distribution and therefore, the singletons should be “strong” enough to beat the effects of any interactions in the joint:

$$P(X) = \prod_{r \in \mathcal{V}} \sum_{\mathcal{M} \in \mathcal{P}(\Omega) \cup \{o\}} P(X_r, \mathcal{M}) \quad (25)$$

Since outliers and models are competing for more likelihood across data points, an optimal joint assignment of labels X_{opt} to the dataset can be thought as the one that minimizes the probability of outliers in the joint. In other words,

$$X_{opt} = \underset{X}{\operatorname{argmin}} \left\{ \prod_{r \in \mathcal{V}} \left(1 - \sum_{\mathcal{M} \in \mathcal{P}(\Omega)} P(X_r, \mathcal{M}) \right) \right\} \quad (26)$$

Let $P(\mathcal{X}_r, m)$ the greatest joint probability over all X_r and $\mathcal{M}$. By factoring out this joint, (26) becomes:

$$\mathcal{X} = \underset{X}{\operatorname{argmax}} \left\{ \prod_{r \in \mathcal{V}} \left[P(\mathcal{X}_r, m) \sum_{\substack{\mathcal{M} \in \mathcal{P}(\Omega) \\ \mathcal{M} \neq m}} \frac{P(X_r, \mathcal{M})}{P(\mathcal{X}_r, m)} \right] \right\} \quad (27)$$

Equation (27) implies that maximizing the global likelihood is practically a problem of finding an optimal subset of Ω and a maximizing the joint assignment for this set of models. Hence, to “push” $P(X)$ to its maximum value, one needs to maximize $P(X_r|\mathcal{M})P(\mathcal{M})$ where $\mathcal{M}$ is a set of model-only labels (i.e., does not contain the outlier label). The latter implies that the likelihood of the joint $P(X, \mathcal{M})$ can be used as a global energy measure:

$$E(X, \mathcal{M}) = - \left[\sum_{r \in \mathcal{V}} \log P(X_r | \mathcal{M}) + |\mathcal{V}| \log P(\mathcal{M}) \right] \quad (28)$$

The prior $P(\mathcal{M})$ is estimated as a Bernoulli product, such that each of the models in $\mathcal{M}$ corresponds to one (of many) combination of $|\mathcal{V}|$ nodes in a group of size equal to the size of support of the respective support, $\mathcal{S}(m)$:

$$P(\mathcal{M}) = \prod_{m \in \mathcal{M}_t} \frac{1}{\binom{|\mathcal{V}|}{|\mathcal{S}(m)|}} \quad (29)$$

where $\binom{n}{k} = \frac{n!}{(n-k)!k!}$. By substituting (29) in (28), we obtain the following expression for the global energy measure:

$$E(X, \mathcal{M}) = - \left[\sum_{r \in \mathcal{V}} \log P(X_r | \mathcal{M}) - |\mathcal{V}| \sum_{m \in \mathcal{M}_t} \log \left(\binom{|\mathcal{V}|}{|\mathcal{S}(m)|} \right) \right] \quad (30)$$

The problem with the energy expression in (30) is that it regards all data points as inliers, whereas the GSW joint contains outliers. To circumvent this “incompatibility”, the smallest model likelihood is used for data points labeled as outliers. This is a valid choice since, although it will be much lower than the GSW un-normalized negative log-likelihood, it will nevertheless be comparable to the energy obtained by a GSW sweep over another set of candidate models.