Department of Electrical and Computer Systems Engineering

Technical Report MECSE-18-2004

Spatially consistent 3D motion segmentation

Konrad Schindler

MONASH UNIVERSITY

Spatially consistent 3D motion segmentation

Konrad Schindler
Insititute for Vision Systems Engineering, Monash University
Wellington Road, Clayton, 3800 Victoria, Australia

December 17, 2004

Abstract

3D motion segmentation is the task to cluster corresponding points in multiple (at least two) images, so that each cluster corresponds to a 3D motion in the underlying 3D scene. The problem can be divided into two stages: first, all motion models required to describe the scene have to be found. Second, each correspondence has to be assigned to the correct model. This report is concerned with the second part. A natural procedure is to assign each correspondence to a motion, such that the a-posteriori likelihood of the description is maximized. However, this is not trivial, since the likelihoods of different correspondences are not independent: neighboring correspondences tend to belong to the same motion, a fact commonly referred to as "smoothness" or "spatial consistency". To account for this fact, we model the set of multiview correspondences as an irregular Markov random field (MRF). The MRF is then optimized with recent graph-based methods, and individual clique potentials are inspected for fine-grained outlier detection.

1 Introduction

Motion segmentation is a frequent task in computer vision: a set of image points tracked through a sequence of images shall be classified such that the motion of the points is consistent within each class. A consistent motion is defined as one, which satisfies a parametric motion model. This model can be a simple 2D motion in the image plane, a 3D motion in the scene, or a more complex motion, such as an articulated model for non-rigid deformations. Motion segmentation is a chicken-and-egg problem: to estimate the models for all present motions from the data, the clustering must be known, but to cluster the input data, the motion models are required. The difficulty can be overcome in different ways, e.g., one can alternate between the two steps in an EM-type procedure [15]. Another possibility is to allow points to contribute to the estimation of multiple motions, apply a robust technique, and in a second step assign each point to one of the motions.

No matter what method is used, at some stage we need a mechanism to assign each point to one of the recovered motions. This mechanism must take into account the "smoothness" of the world, i.e., the intuitive notion that the points belonging to the same motion are also spatially clustered in the image. This fact has been widely acknowledged in the literature on 2D motion segmentation [14, 9, 5, 17]. Recently, however, several authors have considered motion segmentation with 3D motion models (or, equivalently, structure-and-motion of dynamic scenes), and in this context spatial consistency has been largely ignored. Instead, the naive solution is used to assign each point to the model it fits best based purely on residuals [15, 16].

A standard method to model local interactions in labeling problems is the Markov random field model. The contribution of this report is to generalize this method, which is widely used in dense 2D segmentation, to the case of 3D motion segmentation. The set of tracked image points is represented as a first order Markov random field (MRF), and an optimal segmentation is found through energy minimization with graph-cuts. It is also shown that the local behavior of the MRF can be used to find miss-matches, which cannot be found with other means, since they accidentally satisfy the constraints of a weak motion model.

For the rest of the paper, we will not be concerned with finding the motion models required to describe the scene, but will assume, that a set of such models exist, to which we have to assign the correspondences. We will however include a rejection class, to which those correspondences are assigned, which with a high probability are incorrectly matched and do not satisfy any of the models. In section 2 we will briefly recall the Markov random field approach, formulate the task in terms of an MRF and specify its components.

Section 3 shows experimental results of the proposed method, and section 4 summarizes and concludes the paper.

2 Markov random fields

2.1 Basics

Markov random fields (MRF) are a probabilistic way of expressing spatially varying priors, in particular smoothness. They were introduced into computer vision by Geman and Geman [6], and have been applied to a wide variety of problems such as image restoration [1, 4], stereo matching and optical flow estimation [4, 10] or higher-level grouping [7, 8]. A Markov random field consists of a set of sites $\{p_1 \ldots p_n\}$ and a neighborhood system $\{N_1 \ldots N_n\}$, so that N_i is the set of sites, which are neighbors of site p_i. Each site contains a random variable U_i, which can take different values u_i from a set of labels $\{l_1 \ldots l_k\}$. Any labeling $U = \{U_1 = u_1 \ldots U_n = u_n\}$ is a realization of the field. The field is a MRF, if and only if each random variable U_i depends only on the site p_i and its neighbors $p_j \in N_i$. Each combination of neighbors in a neighborhood system is called a *clique* C, and the prior probability of a certain realization of a clique is called the *clique potential* V_C. The basis of practical MRF modeling is the Hammersley-Clifford Theorem, which states that the probability of a realization of the field is related to the sum over all clique potentials via $P(U) \propto \exp(-\sum V_C(U))$. A standard reference for MRFs in computer vision is [11].

If only cliques of 1 or 2 sites are used, the field is called a first order MRF, and

$$P(U) \propto \exp\left(-\sum_{p_i}\sum_{p_j \in N_i} V_{ij}(u_i, u_j)\right) \tag{1}$$

The 1-site clique for each p_i is just the clique itself, with likelihood w_i. Each 2-pixel clique consists of p_i and one of its neighbors, and has the likelihood $p_{ij} = \exp(V_{ij}(u_i, u_j))$. Following Bayes' theorem, the most likely configuration of the field is the one which maximizes the posterior energy function

$$E(U) = \sum_{p_i}\sum_{p_j \in N_i} V_{ij}(u_i, u_j) - \sum_{p_i} \ln(w_i) \tag{2}$$

It remains to define the clique potentials V_{ij}. If the goal is smoothness, and the set of labels does not have an inherent ordering, a natural and simple definition is the *generalized Potts model* [3]

$$V_{ij} = \begin{cases} d_{ij} & \text{if } u_i \neq u_j \\ 0 & \text{else} \end{cases} \tag{3}$$

If two neighboring sites have the same label, the incurred cost is 0, if they have different labels, the cost is some value d_{ij}, independent of what the labels u_i and u_j are. In an irregular MRF, the d_{ij} can be some monotonically decreasing function of the distance between the sites p_i and p_j in order to model decreasing influence of the neighbors with increasing distance.

2.2 Defining cliques

The neighbors of each site are the sites, which shall influence its labeling. In a dense motion field the neighborhood system is naturally given by the pixel raster, whereas in an irregular MRF, there are different possible definitions. The first and simplest one is to define a points' neighbors as all points within a certain radius [11]. However, the definition also has some drawbacks:

- first, it ignores the geometric layout of the neighboring points. At first glance it may seem reasonable that in areas with high point density each point has more neighbors and their influence becomes higher. But this also implies that the interaction between two points is independent of the other points in the region. This contradicts the topological notion of neighborhood: if on the straight connection between points `A` and `B` there is a third point `C`, then `A` and `B` should interact *indirectly* via `C`, rather than have a direct link bypassing `C`.
- secondly, the problem arises to determine the radius of the neighborhood, which may vary greatly depending on the density and distribution of correspondences in the image planes. The problem is aggravated by the fact that in real images the feature density is often highest close to the border of an independently moving object, where a lot of gradient information is present.
- thirdly, the distance-induced definition of neighborhood has the effect that the number and influence of neighbors per point varies greatly, if the points are not evenly distributed in the image. This makes it difficult to find a global scale for the clique potentials.

Both the theoretical limitations and practical experiments suggest that a topologically motivated definition of neighborhood is more suitable. The Delaunay triangulation is a standard algorithm to establish a topology for a 2D point set, which is in a certain sense optimal [12]. A topological relation is established between the points, so that the set of neighbors is locally adapted to the point density, and the number of neighbors is distributed more evenly (the average number of neighbors per point converges to 6 as the number of points $N \to \infty$, and is > 5.5 already for very small meshes).

Motion segmentation requires several (at least two) images, and since parts of the scene are moving relative to each other, the neighborhood system will not be the same in different images. The total neighborhood of a correspondence p_i thus consists of all points which are neighbors of p_i in *any of the images.* If a correspondence p_j is a neighbor of p_i in different images, there is a clique C_{ij} for *each of these images.* The local clique potential for a certain realization of p_i and its neighbors $p_j \in N_i$ is the sum over all C_{ij} in all images, where the same pair $\{i, j\}$ may contribute more than once, and with different d_{ij}, if seen in more than one image.

2.3 Clique potentials

For 3D motion segmentation, the possible labels are the different motion models, and we have to assign each correspondence a label. The 1-site clique potential is the likelihood of a correspondence m_i given the motion model T_j. Given a set of motion models in implicit form, $T_j(x) = 0$, each with a standard deviation σ_j, and a set of N correspondences m_i, we can compute the residual for each correspondence in each motion model as $r_{i,(j)} = T_j(m_i)$. The likelihood of m_i given T_j under the assumption of normally distributed residuals is

$$w_{i,(j)} = -\frac{r_{i,(j)}^2}{2\sigma_j^2} - \frac{1}{2}\ln(2\pi\sigma_j^2) \tag{4}$$

For the 2-site clique potentials the generalized Potts model can be applied, as described above. There are different possible definitions of the d_{ij}, and indeed the most successful applications of MRFs in vision use empirical clique potential functions, e.g. [3]. Obviously, a neighbor's influence should decrease with increasing Euclidean distance E_{ij} between corresponding points, so we can write $d_{ij} = \lambda f(E_{ij})$, where $f(x)$ is a monotonically decreasing function and λ is the parameter which controls the amount of smoothing. It turns out that the choice of the right function is not critical. Functions used in our experiments include

- linear decrease $f = 1/x$

- quadratic decrease $f = 1/x^2$
- sublinear decrease $f = 1/\sqrt{x}$
- exponential decrease $f = exp(-\frac{x}{a})$, $5 < a < 15$
- quadratic exponential decrease $f = exp(-\frac{x^2}{a})$, $50 < a < 250$

With an appropriately set λ, any of the given potential functions performs well. As explained below, we recommend the use of relative distances, so that λ is independent of the point density and can only has to be determined once. In all given examples, we used the same smoothing $\lambda = 100$.

Unfortunately, directly using the given functions to compute the d_{ij} leads to a problem, if the distribution of the points in the image is uneven, which is the case in most practical applications. If distances between neighbors differ strongly between different regions of the image, it is no longer possible to set a meaningful global parameter λ. Low λ will perform satisfactory in dense regions, but will not have any smoothing effect in sparse regions, while high λ will perform well in sparse regions, but oversmooth in dense regions. A solution is to use relative distances. The clique potentials for all neighbors of a point p_i are normalized, such that they add up to a constant. Thus, the total influence of the neighbors on each point's labeling is the same, while the influence is distributed among the neighbors according to the potential function.

$$d_{ij} = \lambda \frac{f(E_{ij})}{\sum\limits_{k \in N_i} f(E_{ik})} \tag{5}$$

Note that with relative weights, the clique potentials are no longer symmetric, $d_{ij} \neq d_{ji}$.

2.4 Optimization

With the given formulation, finding the most likely segmentation is equivalent to minimizing the energy functional (2) over the space of realizations F of the MRF. This is a combinatorial problem, which is NP-hard for >2 labels, but can be exactly solved in low polynomial time for only 2 labels with the min-cut/max-flow algorithm [2]: the MRF is converted into a graph, where the sites p_i are the nodes, and the cliques C_{ij} are the arcs joining the nodes p_i and p_j, with cost V_{ij}. Furthermore the graph is augmented with two *terminal nodes* for the two labels, which are connected to every node of the graph with an arc representing the corresponding likelihood w_i (plus a constant which is larger than the maximum possible clique potential for one node). The minimum cut on this graph partitions it into two sub-graphs, such that each node is only connected to one terminal (label).

Recent work on *multi-way cuts* has extended this method to more than two labels [3]. This method is capable of efficiently finding a strong local minimum through pairwise iteration of two-label cuts. Although theoretically still dependent on the order, the resulting minima are "strong" in the sense that the solution cannot be improved by transferring any subset of a class to another class. A large number of experiments with random starting values show that the result for our problem is independent of the initial solution, and we believe that in the case of sparse 3D motion segmentation, with relatively few points and very few labels, the global minimum is usually found. Independent of the initialization the method needs ≤3 iterations over all labels to converge. Note that when using relative distances to compute the clique potentials, the potentials are not symmetric, and thus the graph is directed (whereas it is undirected in the original multi-way cut formulation). However, the method is still valid, as the underlying min-cut method is valid for directed graphs.

2.5 Dealing with outliers

In any real application, the set of correspondences will contain outliers, which do not correspond to images of the same 3D world point. The reason are imperfections of the underlying method to

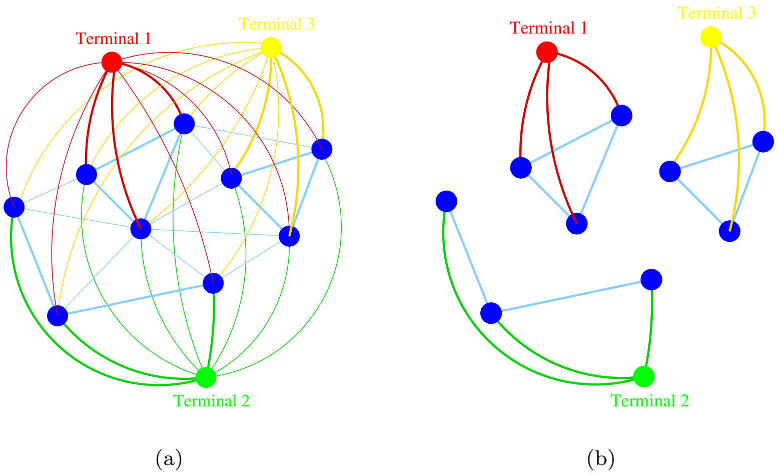


Figure 1: Labeling through graph cuts. (a) Graph representing a MRF before segmentation. Line width denotes edge weight. (b) The result of multi-way cutting.

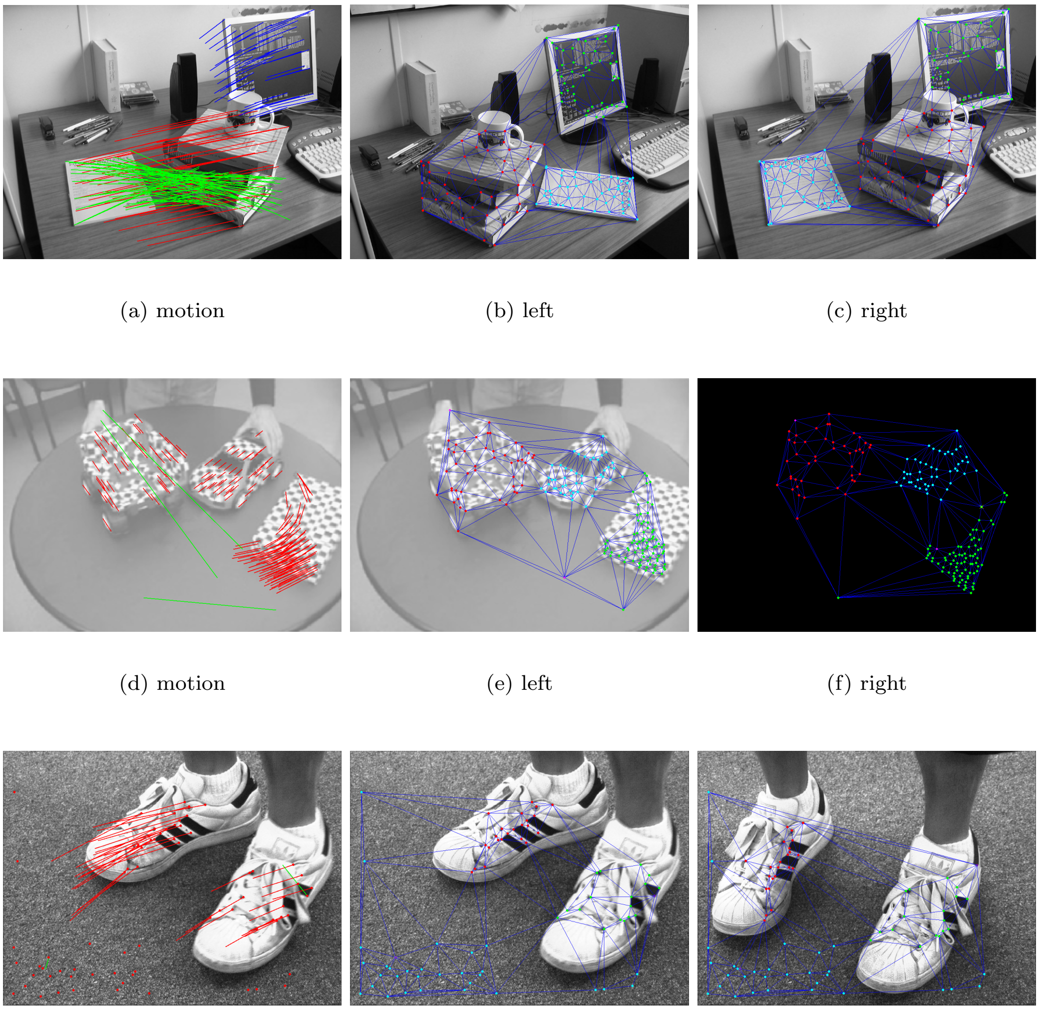
track points or measure optic displacement. For the segmentation to be correct, these points have to be filtered out and assigned an own label, the "rejection class". There are two types of outliers to be considered: the first are points, which do not satisfy any of the models within reasonable bounds (e.g., a multiple of the standard deviation σ). Their likelihood is low in all motion models, and it is trivial to discard them with a threshold on the normalized residual. In most cases this will already have happened during motion model estimation. However, there is a second class of points, which are also miss-matches, but *do* satisfy one of the motion models. This is possible, because the 3D motion models do not enforce spatial consistency, so a correspondence may fulfill the constraints of the model while still being incorrect. Such points are quite frequent especially if only two views are given, because the constraints of some 3D motion models are relatively weak. In particular, the epipolar constraint only requires a point in the second view to lie anywhere on the epipolar line defined by the point in the first view.

In many cases, such points nonetheless violate the assumption of spatial consistency. Therefore, the local properties of the MRF can be used to detect them (assuming that the outliers are not clustered). The described point will have high likelihood in one motion model, but will be surrounded by points from different models. Hence, its relative clique potential will be high (say, $E_{rel} > 70\%$), indicating strong local tension between residual and consistency. With a *threshold on the relative clique-potential* of the converged MRF it is thus possible to detect and remove these points. The threshold is the weighted percentage of neighbors "pulling" the point to other classes, and as such is scale-independent.

3 Experiments

The proposed 3D motion segmentation method was tested with 2-view motion models on several different image pairs. Correspondences were found with the KLT-tracker for the "cars" and "shoes" images, and manually for the "desk" sequence. With a robust multibody structure-and-motion method, a set of fundamental matrices and homographies was automatically recovered, which best explains the correspondences. Details can be found in [13]. Outliers, which do not satisfy any of the recovered motions where removed from the data before segmentation. The labeling was randomly initialized and the described method was applied. The smoothness parameter was set to $\lambda = 100$ for all experiments. The results are shown in Figure 2.

Using a triangulation to define the neighborhood system has an additional advantage: it is

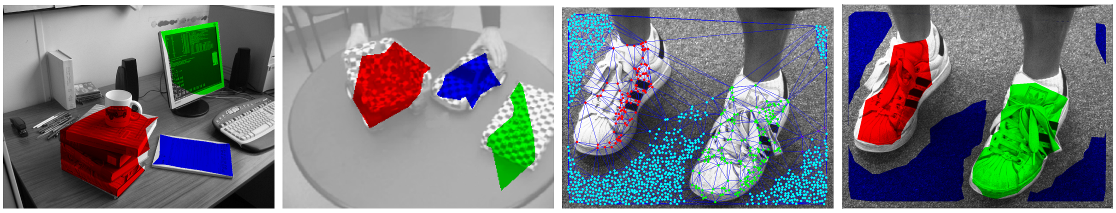


(a) motion (b) left (c) right

(d) motion (e) left (f) right

(g) motion (h) left (i) right

Figure 2: 3D motion segmentation results for 3 different image pairs. The left column shows the image motion overlayed on the first image of the pair. Note that some outliers in the second and third example have survived the model fitting (shown in green). The center and right columns show the two images with their respective triangulations superimposed, and the obtained segmentation. Different colors represent different rigid motions, diamonds are points labeled outliers. See text for details.



(a) desk (b) cars (c) shoes (d) shoes

Figure 3: Coarse dense 3D motion segmentation by exploiting the triangulation. See text for details.

straight-forward to obtain a rough estimate of the dense segmentation. Extending the segmentation (again, under the assumption of smoothness) amounts to finding the triangles, whose corner points all fall into the same class, and completely assign them to the class. If the neighborhood is based on the Euclidean distance and no topology is established, it is not clear how to achieve the same effect. The results are given in Figure 3. For the "shoes" sequence, the original set of correspondences is overly sparse, hence the experiment has been repeated with a denser point cloud, obtained by lowering the threshold of the corner detector.

4 Concluding Remarks

The starting point for this work has been the observation that the smoothness constraint routinely imposed in 2D segmentation tasks should also be used in 3D motion segmentation. A standard tool for this task, the Markov random field formulation, has been adapted to the case of segmenting an irregular set of correspondences, with a different neighborhood system in each image. The method has been tested on several data sets.

Acknowledgments

The source code used to efficiently solve the min-cut problem is provided by Vladimir Kolmogorov on his web-page.

References

[1] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B 48*, pages 259–302, 1986.

[2] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in computer vision. In *Proc. 3rd International Workshop on Energy Minimization Methods in Computer Vision*, 2001.

[3] Y. Boykov, O. Veksler, and R. Zabih. Markov random fields with efficient approximations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, California*, pages 648–655, 1998.

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[5] D. Cremers. A variational framework for image segmentation combining motion estimation and shape regularization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin*, 2003.

[6] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[7] S. Geman, D. Geman, C. Graffigne, and P. Dong. Boundary detection by constrained optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):609–628, 1990.

[8] F. Heitz and P. Bouthemy. Multimodal estimation of discontinuous optical flow using Markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(12):1217–1232, 1993.

[9] Q. Ke and T. Kanade. A robust subspace approach to layer extraction. In *IEEE Workshop on Motion and Video Computing, Orlando, Florida*, 2002.

[10] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *Proc. 7th European Conference on Computer Vision, Copenhagen, Denmark*, 2002.

[11] S. Z. Li. *Markov Random Field Modeling in Image Analysis.* Springer Verlag, second edition, 2001.

[12] O. R. Musin. Properties of the delaunay triangulation. In *Proc. 13th Annual ACM Symposium on Computational Geometry*, pages 424–426, 1997.

[13] K. Schindler. Simultaneous, robust fitting of multiple 3d motion models. Technical Report MECSE-12-2004, Department of Electrical and Computer Systems Engineering, Monash University, 2004.

[14] J. Shi and J. Malik. Motion segmentation and tracking using normalized cuts. In *Proc. 6th International Conference on Computer Vision, Bombay, India*, pages 1154–1160, 1998.

[15] P. H. S. Torr. Geometric motion segmentation and model selection. *Philosophical Transactions of the Royal Society of London A*, 356(1740):1321–1340, 1998.

[16] R. Vidal and S. Sastry. Optimal segmentation of dynamic scenes from two perspective views. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin*, 2003.

[17] J. Wills, S. Agarwal, and S. Belongie. What went where. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin*, pages 37–44, 2003.