International Journal of Computer Vision

DOI: 10.1007/s11263-007-0044-1

Image-Based Modeling by Joint Segmentation

LONG QUAN, JINGDONG WANG, PING TAN AND LU YUAN
The Department of Computer Science and Engineering, The Hong Kong University of Science and Technology
quan@cs.ust.hk
welleast@cs.ust.hk
ptan@cs.ust.hk
luyuan@cs.ust.hk

Received May 10, 2006; Accepted February 14, 2007

Abstract. The paper first traces the image-based modeling back to feature tracking and factorization that have been developed in the group led by Kanade since the eighties. Both feature tracking and factorization have inspired and motivated many important algorithms in structure from motion, 3D reconstruction and modeling. We then revisit the recent quasi-dense approach to structure from motion. The key advantage of the quasi-dense approach is that it not only delivers the structure from motion in a robust manner for practical modeling purposes, but also it provides a cloud of sufficiently dense 3D points that allows the objects to be explicitly modeled. To structure the available 3D points and registered 2D image information, we argue that a joint segmentation of both 3D and 2D is the fundamental stage for the subsequent modeling. We finally propose a probabilistic framework for the joint segmentation. The optimal solution to such a joint segmentation is still generally intractable, but approximate solutions are developed in this paper. These methods are implemented and validated on real data set.

Keywords: structure from motion, image-based modeling, reconstruction, segmentation

1. Introduction

Modeling by video taping has been advocated at CMU by Kanade since the early eighties. The development has been focused on two fundamental achievements by his group, Moravec-Lucas-Tomasi-Kanade feature detection and tracking, and Tomasi-Kanade factorization method for reconstruction.

Lucas-Kanade tracking

The original idea of Lucas and Kanade (1981) is to track or match an image patch $I(x)$ against a reference image $I'(x)$ in two successive frames with unknown transformation $T(\cdot)$ such that $I(T(\mathbf{x})) = I'(\mathbf{x})$. Due to the inherent aperture problem for each individual pixel, it chooses a (weighted) integral error to minimize for the unknown transformation $T(\cdot)$,

$$E(T) = \int_{\mathbf{x}\in W} ||I(T(\mathbf{x})) - I'(\mathbf{x})||^2 (weight)\, d\mathbf{x}.$$

The integral is over a window W around the point $\mathbf{x}$. The transformation T may be taken to be a displacement in its simplest form, $\mathbf{x} \mapsto T(\mathbf{x}) = \mathbf{x} + \mathbf{d}$. Taylor expanding the image term $I(\mathbf{x} + \mathbf{d})$, and differentiating $E(\mathbf{d})$ with respect to $\mathbf{d}$ for minima, we obtain $\mathbf{Hd} = \mathbf{e}$, where $\mathbf{e}$ is the residual error vector $\int_{\mathbf{x}\in W}(I' - I)\mathbf{g}(weight)d\mathbf{x}$ and $\mathbf{g}$ is the gradient of the image I. The $\mathbf{H}$ is the Hessian matrix,

$$H = \int_{\mathbf{x}\in W} \mathbf{g}^T \mathbf{g}(weight)d\mathbf{x}.$$

As done in Tomasi and Kanade (1991) and Shi and Tomasi (1994), the feature point is defined to be the pixel at which we could have a reliable solution, i.e. H is not singular at it.

The Lucas-Kanade tracking equation has several ramifications. First, it leads to one definition of good feature points (Tomasi and Kanade, 1991; Shi and Tomasi, 1994), of which H is well conditioned. This is *per se* the same detector of point of interest or corner in Harris and Stephens (1988) that is motivated by improving Moravec's points of interest (Moravec, 1981), which in turn is from the solution of discretization of the equation $E(T)$ with

$\mathbf{d} = \{(\pm 1 \mid 0, \pm 1 \mid 0)\}$ for self-matching $I = I'$. It is similar to that developed by Förstner (1994) as well. All detectors of point of interest (Förstner, 1994; Harris and Stephens, 1988) only differ at computational implementation of considering the eigensystem of H. Harris and Stephen used $det - kTrace^2 = \lambda_1\lambda_2 - k(\lambda_1 + \lambda_2)^2$. Tomasi and Kanade suggested $\min(\lambda_1, \lambda_2) > \lambda$, and Förstner took aslo the scale σ into account and proposed $\kappa^2 = \frac{\lambda_2}{\lambda_1\sigma^2}$.

Second, it suggests that a dense correspondence between frames is ill-posed in nature as H is not well conditioned everywhere. I.e. there are pixels for which H are close to singularity. The impossibility of dense correspondence justifies the necessity of the introduction of the quasi-dense correspondence (Lhuillier and Quan, 2005) as a more achievable goal. Thirdly, it implies that feature point based approach, now the mainstream, is by its definition—the infinitesimal expansion $\mathbf{d}$ for matching and self-matching—only good for close frames. This indeed reflects the actual practice in the representative systems implemented in Pollefeys et al. (1998) and Nister (2001). Though a more general transformation than the translational displacement could be introduced into the development of matching and self-matching equations without theoretical difficulties (Triggs, 2004), it leads often to inconclusive tradeoff between invariance and rareness of the descriptors (Triggs, 2004).

Tomasi-Kanade factorization algorithm

Related to the tracking, given the tracked points over the sequence, but one step further to reconstruct these feature points in 3D space. By first considering a simplified camera projection model that is approximated by an orthogonal type projection model $(x, y, z) \mapsto \lambda(u, v)$ instead of a more general central projection model. Tomasi (1991) proposed the factorization algorithm for reconstruction. Given n points tracked over m views, and stack all measurements (u, v) over all views to form a big measurement matrix W, which has a rank constraint by its construction $W = MS$. Using SVD on M results in $W = U\Sigma V = (U\sigma^{1/2})(\sigma^{1/2}V) = MS$. The motion M and shape S are defined up to an arbitrary affine transformation A as $W = MS = (MA^{-1})(AS)$, so the M and S are *de facto* the affine motion and affine shape. Next, the metric constraints to guarantee that M is a valid rotation matrix are used to finalize the Euclidean motion M and Euclidean shape S. Each of these two steps contains fundamental concepts that are related to the development of the uncalibrated approach.

First, the importance of the pre-metric structure of the shape, the S resulted from the SVD, though not elucidated, but was related to the concept of affine shape introduced by Koenderink and Van Doorn in 1988. Koenderink's paper (Koenderink and van Doorn, 1989) has already been in circulation in 1988, and was published in 1991 (Koenderink and van Doorn, 1989). The technical report of Tomas-Kanade factorization (Tomasi, 1991) was available in 1991 and its journal version (Tomasi and Kanade, 1992) in 1992. This affine shape was innovative, but limited to only affine cameras. It is Faugeras' paper (Faugeras, 1992) that gave a definite answer to what is the parallel to an affine camera for a projective camera. Hartley (1992) independently reached the same results. Mohr et al. (1992, 1995) followed up by proposing a numerical scheme, a kind of bundle adjustment for projective reconstruction for multiple views, using an explicit projective basis defined by 5 space points. Interestingly and importantly, it was shown in Sturm and Triggs (1996) and Triggs (1996) that the original Tomasi-Kanade affine factorization for affine cameras could be extended to projective reconstruction for projective cameras, this is the projective factorization, that is the only alternative to a bundle-adjustment type projective reconstruction. It has even an advantage of straightforward initialization when an iterative scheme is necessary. This projective tour for structure from motion is important as it tells us an important fact that the matching problem for a rigid scene or object is encoded by this uncalibrated projective geometry. And this projective structure could be efficiently computed from the feature points detected and matched in multiple views.

Second, it has been known that the second step of the enforcement of metric constraints is capital, but hard to succeed in some cases. This has been primarily viewed as a successive step of the entire reconstruction for affine cameras. Now from an uncalibrated projective viewpoint, the enforcement of the metric constraints is equivalent to the autocalibration of the uncalibrated approach. Indeed, using the orthogonality constraints originally proposed by Tomasi and Kanade gives an alternative way of formulating the autocalibration both for affine and projective cameras, which is usually formulated in terms of the absolute conic borrowed from the projective geometry.

Combining Lucas-Kanade-Tomasi feature detection and tracking with the projective factorization is an alternative of the standard uncalibrated approach which is often implemented (Pollefeys et al., 1998; Hartley and Zisserman, 2000; Faugeras et al., 2001) by first extracting the points of interest, Harris corners, then combining with robust statistics such as RANSAC or LMS with a projective structure (Zhang et al., 1995), and finally optimizing the structure by a bundle-adjustment-like optimization (Mohr et al., 1995, 1992; Triggs et al., 2000) in projective space.

Figure 1. One overview on the right of the reconstructed quasi-dense points for the entire scene from 25 images shown on the left.

Quasi-dense approach

This sparse structure from motion approach usually requires a dense frame rate and leads to a too sparse set of points to be sufficient for object modeling and depiction. This insufficiency motivated the development the quasi-dense approach started since 1998 in Lhuillier (1998) and matured into the work in Lhuillier and Quan (2005). It is robust and handles more distant views. More importantly, it produces a set of semi-dense 3D points that was impossible with the previous methods. The quasi-dense approach will be briefly revisited in Section 2. One example is shown in Fig. 1. The quasi-dense approach can be viewed as an extension to the discussed two key subjects, feature point selection, re-sampled quasi-dense point vs. interest point, and reconstruction, hierarchical division combined with a bundle-like method vs. batch solution combined with a factorization.

Joint segmentation approach

The increased density of the reconstructed 3D points from multiple views, paves the way for the three-dimensional modeling of the objects in space, in addition to the recovery of the camera positions. But the 3D points, even semi-dense, are unstructured in space, therefore are not yet sufficient for creating a geometric model of the underlying objects. It is necessary to group the points and pixels that belong to the same *object* into the same cluster of points and pixels. Obviously, the concept of object is subjective: for the example scenario shown in Fig. 2, the whole plant might be considered as an object for some applications, whereas each individual leaf should be considered as an independent object for some others, depending on the application and the realism details of the modeling required. The main contribution of this paper is to introduce a joint segmentation framework, for both 3D points and 2D pixels, and look for robust and efficient solutions to it in Section 3. Figure 2 shows one of the original 25 images captured by a handheld camera for the example scenario, a rendered image of the final modeling result based on the desired 3D segmentation and 2D segmentation results by a semi-automatic approach developed in Quan et al. (2006) using the joint segmentation method presented in this paper. The segmentation and modeling of such complex objects are almost impossible without the joint segmentation.

2. Quasi-Dense Approach Revisited

The quasi-dense approach developed in Lhuillier (1998); Lhuillier and Quan (2002, 2005) overcomes the sparseness of feature points and results in a more efficient and robust algorithm when combined with a bundle adjustment like algorithms both at projective and Euclidean stages. The purpose is to view the quasi-dense approach as an extension of the discussed two key subjects: feature point, quasi-dense point vs. interest point; and reconstruction, hierarchical division combined with a bundle-like method vs. batch solution with a factorization scheme.

2.1. Quasi-Dense Point

Initialization by sparse points of interest. We start by detecting the points of interest in each image (Zhang et al., 1995), then compute the Zero-Mean Normalized Cross-Correlation (ZNCC) to match points of interest in two images. First, we do the correlation from the first image

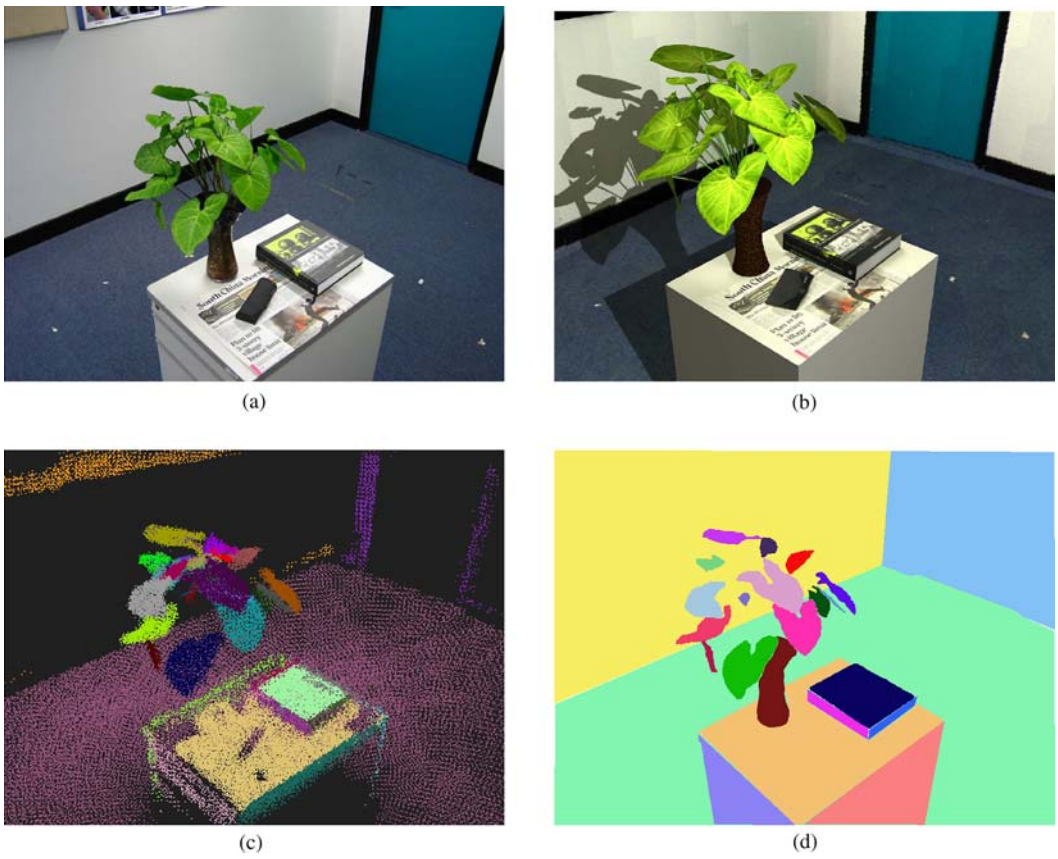


Figure 2. A proper segmentation is fundamental to 3D modeling. (a) One of the 25 original images captured by a handheld camera. (b) A rendered image at a similar viewpoint of the reconstructed 3D model based on the segmentation results in (c) and (d). (c) A desired segmentation of 3D data points from the reconstructed quasi-dense points, and (d) a desired segmentation of an input image.

to the second, and then do the same from the second to the first to retain a one-to-one consistent matching of points of interest between two images. Most of the standard approaches will continue with this set of feature points by introducing global geometry constraint. The quasi-dense approach will not, it is merely initialized at this stage.

Propagation. The initial sparse matches of points of interest, still often containing a significant portion of outliers, are now sorted by decreasing ZNCC correlation score. These sorted seed points bootstrap a region-growing type algorithm that propagates the matches locally from the most reliable pixels to the less reliable ones. At each step, the match $(\mathbf{x}, \mathbf{x}')$ with the highest ZNCC score from the current list of the seeds is taken off the list and is propagated in the immediate spatial neighborhood $N(\mathbf{x}, \mathbf{x}')$ for potential new matches. The new match has to be sufficiently discriminant and satisfy the disparity gradient constraint for uniqueness. Each new match is revaluated by the ZNCC to update the sorted list of the seeds for the next step. In addition to the important the best-first propagation strategy, there are two key features of this approach.

First, the key is the definition of the potential match candidates $N(\mathbf{x}, \mathbf{x}')$ satisfying the disparity gradient (Lhuillier and Quan, 2002; Pollard et al., 1985) as

$$N(\mathbf{a}, \mathbf{a}') = \{(\mathbf{b}, \mathbf{b}'), \mathbf{b} \in N(\mathbf{a}), \mathbf{b}' \in N(\mathbf{a}'), \\ | \ (\mathbf{b}' - \mathbf{b}) - (\mathbf{a}' - \mathbf{a}) \ | < \epsilon\},$$

given the neighborhood $N(\mathbf{x})$ consisting of all pixels of $\mathbf{x}$ within a window of size $n \times n$. Figure 3 illustrates one

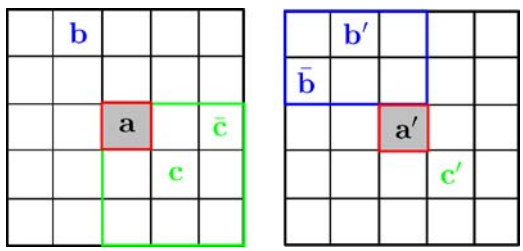


Figure 3. Possible matches $N(\mathbf{a}, \mathbf{a}')$ around a seed match $(\mathbf{a}, \mathbf{a}')$ come from its 5×5-neighborhood $N_5(\mathbf{a})$ and $N_5(\mathbf{a}')$. The possible matches for $\mathbf{b}$ and $\mathbf{c}'$ are in the 3×3 window frame centered at $\mathbf{b}'$ (resp. $\mathbf{c}$).

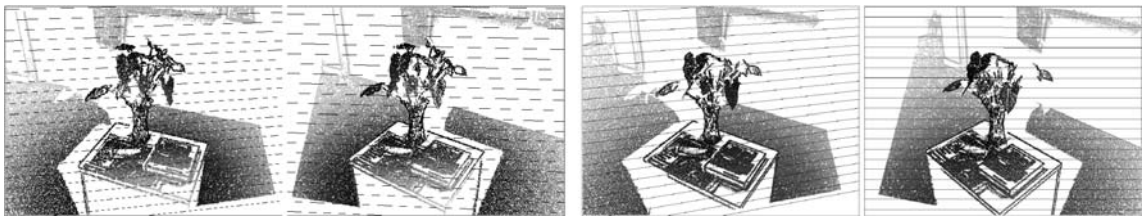


Figure 4. The quasi-dense disparities for the two pairs of views (f_1, f_2) and (f_2, f_3) of a triplet of views (f_1, f_2, f_3).

example of possible matches generated from $(\mathbf{a}, \mathbf{a}')$ for $n = 5$ and $\epsilon = 1$. The topologies of the matches might be very different in different images: $(\mathbf{a}, \mathbf{a}')$, $(\mathbf{b}, \bar{\mathbf{b}})$, $(\bar{\mathbf{c}}, \mathbf{c}')$).

Second, the sufficient density is achieved for largely separated views because of the confidence measure used in the approach, it is defined in Lhuillier and Quan (2005) as $\max\{E\}$, where $\{E = |I(\mathbf{x} + \mathbf{d}) - I(\mathbf{x})|, \mathbf{d} \in \{(\pm 1, 0), (0, \pm 1)\}$. It simply responds to a gradient in any direction, and is less restrictive than Moravec operator (Moravec, 1979) that is the local maximum of $\min\{E, \mathbf{d} \in \{(\pm 1, 0)\}$ that again approximates the intended summation $\sum_i E_i^2$ in the least squares matching. It therefore allows the matches to also go along edge points despite the aperture problem while avoiding matching uniform areas by imposing a threshold on the confidence. An example of quasi-dense disparities is shown in Fig. 4.

Quasi-dense points as re-sampled correspondence points. Disparities give pixel-to-pixel correspondences, which are locally dense, but they give an irregular distribution of clusters of pixels, which is not suitable for geometry computation. Clustered pixels bias the geometric constraints and increase the computation cost. We re-sample the disparities to produce the quasi-dense points that are more suitable for subsequent computations. Re-sampling not only produces a reduced set of point correspondences that are more uniformly distributed than the raw pixel correspondences, moreover, it acts as a post-propagation regularization that improves the reliability of the estimation of the disparties by integrating local geometric constraints. The scene or object surfaces are often at least locally smooth, and this local smoothness is encoded by a local plane homography.

The first image plane is partitioned into a regular square grid of $n \times n$ pixels. This size n is a trade-off between the sampling resolution and regularization stability. For each square patch, all pixel correspondences inside it induced by the current disparities are used to fit a plane transformation. The transformed points, in subpixel accuracy, are considered as the quasi-dense correspondence points. An example of quasi-dense disparties is shown in Fig. 4, and an example of quasi-dense points is shown in Fig. 5.

2.2. *Quasi-Dense Geometry Reconstruction*

Projective quasi-dense geometry. We consider a linear sequence of views that requires sufficient overlapping between every two adjacent views. We essentially use the hierarchical structuring of the subsequences adapted in Laveau (1996). A sequence of images indexed by $[i, \ldots, j]$ is recursively partitioned into two subsequences $[i, \ldots, k, k+1]$ and $[k, k+1, \ldots, j]$ with the two overlapping frames k and $k+1$, where k is the median of the index range $[i..j]$. The partition stops at each triplet of views for which the geometry is computable and estimated. Then the geometry of each two subsequences are merged by computing a space homography induced by the the two overlapping frames of two subsequences. The geometry of the merged sequence is always re-optimized as a whole. The geometry of a triplet of views is first

Figure 5. The resulting quasi-dense correspondence points for a triplet of views. The points in red are outliers.

directly computed using 6 corresponding points in three views (Quan, 1995), similar to 7 points in two views, then optimized subsequently.

The establishment of quasi-point correspondences is to divide the three views into two pairs with one overlapping view. The quasi-dense point correspondences for a pair of views discussed in the previous section are now used to compute an initial solution of fundamental matrix of the pair. To be more robust and accurate, a second *constrained propagation* by the current fundamental matrix is re-computed and resampled to obtain the final quasi-dense correspondences of the pair. The quasi-dense point correspondences for each pair is propagated into the triplet thanks to the common view.

It is important to observe that the two propagations improve the robustness of the correspondence computation. The first unconstrained propagation has the advantage of overcoming the biased estimates toward the areas with a high density of matches usually observed in other feature based approaches due to the irregular distribution of the points in image space, discussed as well in Hartley and Zisserman (2000).

Euclidean quasi-dense geometry. The projective geometry of 3D quasi-dense points and cameras is upgraded into Euclidean structure by an initial autocalibration followed by an optimization in Euclidian space, a bundle adjustment. The initial value for the focal length might be either computed by some linear autocalibration methods (Nister, 2001; Triggs, 1997) or obtained from the setting of the digital camera. The autocalibrated intrinsic parameters transform the reconstruction into a metric representation. The world Euclidean coordinate frame is fixed at the camera center of the view in the middle of the whole sequence and the scale is fixed to be the maximum distance between any pair of camera center positions. We re-parameterize each calibrated camera by its six individual extrinsic parameters and one intrinsic focal length to finally adjust the geometry of the whole system through the optimization. This natural re-parametrization treats equally all cameras for uncertainty estimation, but leaves the seven d.o.f scaled Euclidean transformation as the gauge freedom (Lhuillier and Quan, 2005; Triggs et al., 2000).

2.3. Remarks

The quasi-dense approach is remarkably both more robust and accurate than the standard approaches available in Hartley and Zisserman (2000), Faugeras et al. (2001), Pollefeys et al. (1998) and Nister (2001). For the example scenario in Fig. 2 the reconstructed 3D quasi-dense points are shown in Fig. 1. Some quantitative comparisons are presented in Lhuillier and Quan (2005). We support the previous analysis by mentioning the two keypoints for the achievements of robustness and accuracy: wide separation for adjacent views and density of the corresponding points. It indeed works for more widely separated image pairs than that required by the standard sparse approach. It is due to the fact that the number of matched interest points drastically decreases with increasing geometric distortion between views in the feature point based approach. The propagation strategy used for the quasi-dense points accepts the more complicated local image distortion because of the 2D disparity gradient constraint developed in Lhuillier and Quan (2002). The more specific wide baseline stereo such as Tuytelaars and Van Gool (2000), Urban et al. (2002) and Lowe (2004) produces far fewer points, which, again, might be sufficient for the camera geometry and indexation, but not for object modeling.

3. Joint Segmentation

Given a set of 3D points, the quasi-dense points in our case such as the example shown in Fig. 1, and all the images of the sequence, we would like to segment jointly all 3D points and 2D pixels into groups of meaningful objects.

3.1. Related Work

Image segmentation has been a traditional and fundamental topic in computer vision. There is an abundant literature in both segmentation of 3D range images or depth maps, and that of normal 2D gray level or color images. Segmentation of range images usually looks for more local geometric characterization as it has a high density of 3D points, it leads to different approaches as it has no associated 2D images. Segmentation of normal images uses only pure pixel information, and attracts recent attentions. Unfortunately it operates only in 2D space with pure pixel information, and its purposes are often motivated by object recognition. Some representative works include (Shi and Malik, 2000; Tu and Zhu, 2002). It is natural that segmentation of multi-view, often calibrated offline, handles both image and depth information. There are several representative works. The layered approaches originated from Wang and Adelson (1994) usually do not directly adopt 3D reconstruction information. In Wills et al. (2003) and Xiao and Shah (2004), motion estimation and segmentation on the extracted correspondences between frames are performed, then layer assignment (i.e. pixel label) is obtained through propagating the labels of the corresponding pixels. In Patras et al. (2001), joint inference of motion estimation and labeling is solved using the Expectation Maximization (EM) algorithm. The modern stereo matching approach is much similar to the

Figure 6. One source image is shown in (a). The zoomed subimage is shown in (b). Assume that the Euclidean distances satisfy $|s_a - s_b| > |s_b - s_c|$ in 3D space. The similarity of the colors for a and b makes them closer in grouping, while the presence of the edge inbetween b and c moves them apart in grouping. A typical example in which the leaves can not be separated with 3D Euclidean distances is shown in (c), but can be separated after considering color and contour information as shown in (d).

layered approach, and it in essence discretizes the 3D space into a few layers . One of the most representative works is Kolmogorov and Zabih (2002). Bilayer segmentation as a simplest layered representation, basically uses the learned appearance model instead of correspondence relations. Stereo cues are probabilistically fused for segmentation in Kolmogorov et al. (2005). The edges of the background of last frame are used to attenuate the edges of current frame in Sun et al. (2006). The probabilistic motion model is learnt to help segmentation in Criminisi et al. (2006).

The existing approach may directly integrate the motion constraints into the segmentation, but is often primarily preoccupied by the recovery of depth information rather than the clustering of the objects. It makes little exploitation of reliable 3D information, probably lack of it. Intuitively, there is much richer information available to explore for object segmentation when both 3D and 2D information is available. For instance, some objects are obviously separable in 3D whereas others are clearly cut out by image boundary information even if they are closely connected in space as shown in Fig. 6. We will first formulate the joint segmentation in a probabilistic framework, then propose some solutions to it.

3.2. *Probabilistic Formulation*

Let $I = \{I_i\}$ be the set of n images with $i = 1, \ldots, n$. Each image I_i is represented by a set of pixels in RGB space, i.e. $I_i = \{(\mathbf{u}_k, \mathbf{c}_k)\}$ with k up to the number set by the image resolution, and each image point includes its position $\mathbf{u}_k$ in image space and three colors $\mathbf{c}_k$. It is assumed that all the images are fully calibrated with respect to a common coordinate fame. We define a *joint point* $\mathbf{x}$ to be a vector composed of the 3D coordinates (x, y, z) of a point in space and all its corresponding image points $\mathbf{u}_i$ in all images, i.e. $\mathbf{x} = ((x, y, z), (\mathbf{u}_1, \mathbf{c}_1), \ldots, (\mathbf{u}_n, \mathbf{c}_n))$, where each image point satisfies $\mathbf{u}_i = \mathbf{P}_i(x, y, z, 1)^T$ for the projection matrix $\mathbf{P}_i$ of the i-th camera. The correspondence information is encoded in the joint point representation. And each joint point $\mathbf{x}$ is associated with a n-dimensional visibility vector $\mathbf{v}$ with binary values to indicate that $\mathbf{u}_i$ is visible in the i-th image if the i-th component is 1, and invisible otherwise. A segmentation is a set of labels $L = \{l_k\}$, and each of them l_k assigns a set of joint points to a common group. The number of the labels is unknown. If $X = \{\mathbf{x}_j\}$ is the given set of joint points, and $V = \{\mathbf{v}_j\}$ the given set of visibilities for each joint point, X and V are given by the quasi-dense reconstruction in our case, then L_X is a labeling for the given set of joint points X. The inference of L_X could be treated as a maximum *a posteriori* estimate of the probability $P(L_X \mid X, V, I)$. A joint segmentation L_X is formally given by

$$L_X = \arg\max_{L_X} P(L_X \mid X, V, I). \tag{1}$$

This MAP could be solved by representing the posterior probability $P(L_X \mid X, V, I)$ as a Conditional Random Field (CRF) as in Lafferty et al. (2001),

$$P(L_X \mid X, V, I) \propto \exp -\{E^l(L_X; X, V, I) + \lambda E^s(L_X; X, V, I)\},$$

where $E^l(\cdot)$ is the energy for the likelihood model of the labeling, and $E^s(\cdot)$ is the energy for the conditional prior model of the labeling.

3.2.1. Joint Likelihood Model. For a joint point, there is no evidence to support the dependency of the 3D position $\mathbf{x}^s = (x, y, z)$ and the image colors $\mathbf{x}^c = (\mathbf{c}_1, \ldots, \mathbf{c}_n)$, the joint likelihood model is then divided into the independent spatial and color terms:

$$E^l(L_X; X, V) = \rho_{l_1} E^l(L_X; X_C, V) + \rho_{l_2} E^l(L_X; X_S), \tag{2}$$

where $X_C = \{\mathbf{x}_j^c\}$ is the set of the components of the image colors of the joint points X and $X_S = \{\mathbf{x}_j^s\}$ is the set of the components of 3D coordinates of the joint points X.

Color likelihood. The Color likelihood is taken to be

$$E^l(L_X; X_C, V) = -\sum_{j=1}^{m} \log p\left(\mathbf{x}_j^c \mid l_j\right), \tag{3}$$

where the color likelihood model for each joint point is defined by $p(\mathbf{x}^c \mid l) = \prod p(\mathbf{c}_i \mid l)^{\frac{\mathbf{v}_i}{|\mathbf{v}|_1}}$ as the geometric mean of the likelihoods of all its 2D pixels. The L_1-norm $|\mathbf{v}|_1$ of $\mathbf{v}$ is just the number of visible pixels of the joint point $\mathbf{x}$. The invisible features $\mathbf{c}_i$, with $\mathbf{v}_i = 0$, do not contribute to the model. In other words, the definition actually only performs geometric mean on the visible 2D pixels, which equally treats every joint point without over-using it.

The probability density $p(\mathbf{c} \mid l)$ describes the color distribution of the joint points for the label l. It is natural to take a Gaussian Mixture Model (GMM) (Rother et al., 2004; Blake et al., 2004; Kolmogorov et al., 2005), whose parameters could be estimated by the Expectation Maximization (EM) algorithm (Dempster et al., 1977). Again the visible color of each joint point $\mathbf{x}$ is weighted by $\frac{1}{|\mathbf{v}|_1}$ for the estimation to view equally each joint point.

Spatial shape likelihood. The spatial shape likelihood is used to model the shape prior of the object. It is desirable to be able to model a general surface patch that will be good for modeling the visible surface of an object, but it is usually computationally difficult. We will take a planar model to approximate the shape model of the objects, this is similar to Torr et al. (2001) that uses a plane as a layer. Using a planar model is however insufficient since the 3D points for an object spatially spread rather in a compact manner. To take the compactness into account, we use the Mixture of Probabilistic Principal Component Analysis (MPPCA) (Tipping and Bishop, 1999) to obtain a probability measure. And this shape likelihood term is defined as:

$$E^l(L_X; X_S) = -\sum_{j=1}^{m} \log p\left(\mathbf{x}_j^s \mid l_j\right), \tag{4}$$

where $p(\mathbf{x}^s \mid l)$ is formulated in terms of Probabilistic Principal Component Model (PPCA) (Tipping and Bishop, 1999). The principal 2D subspace $\mathcal{P}_2$ and the complementary 1D subspace $\mathcal{P}_1$ are first computed through Principal Component Analysis (PCA) for the points with the same label. Then the probability is the combination of the two terms. The first term is to penalize the projection in the complementary subspace $\mathcal{P}_1$, i.e. the residual error; the smaller the probability is, the larger the residual error is. The second term is to measure the compactness of the projected points in the principal plane $\mathcal{P}_2$ through a Gaussian model. We assign a larger weight to the first term than the second, slightly different from Tipping and Bishop (1999).

3.2.2. Joint Prior Model. Conditional prior model is used to measure the consistency between the labelings of the joint points. The k-way consistency is usually more powerful to describe the prior than the pair-wise consistency, but it leads to higher computation cost. We adopted the common pair-wise affinity as the prior that is formulated as the following:

$$\begin{aligned} E^s(L_X; X, V, I) &= \sum_{(i,j)\in E} E^s(l_i, l_j; X, V, I) \\ &= -\sum_{(i,j)\in E} a(i, j)\delta(l_i = l_j), \end{aligned}$$

where $a(i, j)$ is the affinity function between the points i and j, E is the set of all pairs of points corresponding to linked edges in a graph, and $\delta(a = b)$ is a Dirichlet function such that $\delta(a = b)$ is 1 if $a = b$ holds and 0 otherwise. The quality of a segmentation based on the formulation fundamentally depends on the affinity, we seek therefore to define it jointly from both 3D and 2D features.

3D Affinity. Closer points in space tend to have higher probability of belonging to the same group, i.e. the distance between the points of the same group is smaller than that of the points of different groups. We naturally take this spatial distance as an affinity measure $a_{3d}(i, j) = \exp(-\frac{||\mathbf{s}_i - \mathbf{s}_j||^2}{2\sigma_{3d}^2})$, where $\sigma_{3d} = \mathrm{E}^{1/2}(||\mathbf{s}_i - \mathbf{s}_j||^2)$. The Gaussian function has the desired properties for an affinity, and is popular in spectral clustering and normalized cut (Shi and Malik, 2000). In addition to the 3D Euclidean distance, the normal directions are also important for shape smoothness. We incorporate the difference between normal directions into the affinity and define $a_{3n}(i, j) = \exp(-\frac{||\mathbf{n}_i - \mathbf{n}_j||^2}{2\sigma_{3n}^2})$, where n_j is the normal direction vector of the point j, approximately estimated from its neighbor points, and $\sigma_{3n} = \mathrm{E}^{1/2}(||\mathbf{n}_i - \mathbf{n}_j||^2)$. The final 3D affinity is given by $a_3(i, j) = a_{3d}(i, j)\, a_{3n}(i, j)$.

2D Affinity. Since a joint point $\mathbf{x}$ is associated with the image colors, we can define an affinity function encoding the color differences as $a_c(i, j) = \exp(-\frac{||\mathrm{E}(\mathbf{c}_i) - \mathrm{E}(\mathbf{c}_j)||^2}{2\sigma_c^2})$, where $\sigma_c = \mathrm{E}^{1/2}(||\mathrm{E}(\mathbf{c}_i) - \mathrm{E}(\mathbf{c}_j)||^2)$, and $\mathrm{E}(\mathbf{c}) = \frac{1}{|\mathbf{v}|_1}\sum_{i=1}^{n} \mathbf{c}_i$. This color consistency between joint points is intuitively estimated using their average colors, since different points may have different numbers of visible color features. Averaging the colors leads to a more stable solution. However, this affinity function only makes

sense between the objects with apparent different colors. In case of apparent similar colors, image contour features, similar to Malik et al. (2001), should be incorporated into the affinity as illustrated in Fig. 6.

It is assumed at present that each pixel $\mathbf{u}$ in view I_v is associated with a response $g_v(\mathbf{u})$ to show the degree of the pixel lying on a contour point. The endpoints of the edge $(i, j) \in \mathcal{E}$ must be both visible at least in one view, meaning that the line segment $[i, j]$ must correspond to a line segment visible in the same view. We can use the following affinity measurement

$$a_{ic}(i, j) = \exp\left(- \frac{\text{med}_v\{\max_{t_v \in [i,j]_v} g_v(t_v)\}}{2\sigma_{ic}^2} \right), \quad (5)$$

where the inner term $\max_{t_v \in [i,j]_v} g_v(t_v)$ finds the maximum contour response along the projected line segment $[i, j]_v$ in view v, the outer term $\text{med}_v\{\cdot\}$ tries to seek the median contour response in all possible views, and σ_{ic} is the variance of the median contour responses of all line segments. Different from Quan et al. (2006), the median operator is observed in experiment to be more robust than the maximum operator.

An edge map in Malik et al. (2001) is used to compute the contour responses in this paper. The oriented filter bank based on rotated copies of a Gaussian derivative and its Hilbert transform are used. Let $f_1(x, y) = G''_{\sigma_1}(y)G_{\sigma_2}(x)$, and $f_2(x, y) = H(f_1(x, y))$ is the Hilbert transform of $f_1(x, y)$ along the y axis. The oriented energy at angle 0° is defined as $E_{0^\circ} = (I * f_1)^2 + (I * f_2)^2$. Then the contour response is defined as $g(x, y) = \max_\theta E_\theta(x, y)$.

Finally, we are able to perform simple multiplication of the affinities to define the joint affinity to be $a(i, j) = a_3(i, j) \times a_c(i, j) \times a_{ic}(i, j)$.

3.2.3. Graph Construction. The set of edges E in defining the likelihood can be constructed using k-Nearest Neighbor (k-NN) technique. As mentioned above, we expect that the points i and j of $(i, j) \in E$ must be both visible at least in one view. This can be guaranteed as follows. Each view is associated with a set of joint points that are visible in this view. We first build for each view a k-NN network on the corresponding set of joint points according to the 3D Euclidean distance, and set k to be 5 by default. Then we combine those networks together to reach a graph on the entire joint points. Finally we discard some incident edges with larger distance for each joint point such that the graph is not so dense for efficient computation. It should be noted that the graph construction is critical in that the 2D affinity definition is entirely based on the construction.

3.2.4. Computation. The final objective function for the joint segmentation is given by

$$E = -\sum_{i=1}^{m} \left(\rho_1 \log p\left(\mathbf{x}_i^c \mid l_i\right) + \rho_2 \log p(\mathbf{x}_i^s \mid l_i) \right) - \lambda \sum_{(i,j)\in\mathcal{E}} a(i, j)\delta(l_i = l_j).$$

This problem is, generally speaking, a graph partitioning problem, but the optimization depends on how discriminative the likelihood is.

In the case of separating objects with apparent color dissimilarity, the color likelihood is usually discriminative, but not the shape model. The objective function reduces to the common combinatorial formulation:

$$E = -\rho_1 \sum_{i=1}^{N} \log p\left(\mathbf{x}_i^c \mid l_i\right) - \lambda \sum_{(i,j)\in\mathcal{E}} a(i, j)\delta(l_i = l_j). \quad (6)$$

This equation is a typical discrete CRF formulation, which can be efficiently optimized using graph cut algorithm (Boykov et al., 2001). Its complexity is of $O(N^3)$ because the graph is sparse, and there are N points and $O(N)$ edges.

In other cases there exists discriminative spatial shape feature, but not discriminative colors. For instance, we may want to model individual leaves, which requires an individual leaf segmentation. The color is similar for all leaves of the same plant, but each leaf lies on a different surface patch in space. The objective function can be simplified to

$$\begin{aligned} E &= -\rho_2 \sum_{i=1}^{m} \log p\left(\mathbf{x}_i^s \mid l_i\right) - \lambda \sum_{(i,j)\in\mathcal{E}} a(i, j)\delta(l_i = l_j) \\ &= -\rho_2 \sum_{i=1}^{m} \log p\left(\mathbf{x}_i^s \mid l_i\right) + \lambda \sum_{(i,j)\in\mathcal{E}} a(i, j)(1 - \delta(l_i = l_j)) + \text{const.} \end{aligned} \quad (7)$$

Now that the problem, similar to those used in Zabih and Kolmogorov (2004) and Zhu and Lafferty (2005), can be casted into an unsupervised clustering framework. We can use an iterative algorithm that combines GMM and graph cut. We rewrite the affinity part in matrix form,

$$\sum_{(i,j)\in E} a(i, j)(1 - \delta(l_i = l_j)) = \text{Trace}(\mathbf{X}^T(\mathbf{D} - \mathbf{W})\mathbf{X}), \quad (8)$$

where $\mathbf{W}$ is the symmetric affinity matrix with each entry $\mathbf{W}_{ij}$ corresponding to the affinity value of the edge (i, j),

and $\mathbf{W}_{ij} = 0$ if there is no edge between the points i and j, $\mathbf{D}$ is the diagonal degree matrix in which the diagonal entry $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$, $\mathbf{X}$ is a label matrix of size $n \times c$ such that $\mathbf{X}_{il} = 1$ if the label of point i is l and 0 otherwise. The minimization of the above function is in essence a graph min-cut problem. As pointed out in Shi and Malik (2000), a normalized cut criterion will give a better graph partition. We use the normalized cut method to get a reliable labeling (graph partitioning):

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \text{Trace}(\mathbf{X}^T \mathbf{D}^{-\frac{1}{2}}(\mathbf{D} - \mathbf{W})\mathbf{D}^{-\frac{1}{2}}\mathbf{X}). \quad (9)$$

We optimize Eq. (7) in two steps. we first use the technique in Shi and Malik (2000) to solve Eq. (9) to get an initial label $\hat{\mathbf{X}}$. In the second step, the PPCA model can be estimated according to each group of points. Afterwards the MPPCA is used to refine the labeling. Although an iterative algorithm can be used for optimization, its convergence can not be theoretically guaranteed and it does not improve significantly the results. We observed in our experiments that the two steps without iteration work satisfactorily, thanks to the affinity function. The most computationally expensive step is sparse SVD for $\mathbf{L} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{D} - \mathbf{W})\mathbf{D}^{-\frac{1}{2}}$ in Eq. (9). Its complexity is of $O(N^2)$ as $\mathbf{L}$ has only $O(N)$ nonzero entries.

4. Segmentation Propagation

The joint segmentation L_X is so far defined for a given set of joint points X, which is, practically, the set of quasi-dense points. Ideally had we a dense reconstruction from multiple views, the joint segmentation could have been applied to the set of dense joint points. This only increases the computational cost. As we have argued at the beginning of the paper that the full dense reconstruction is hardly achievable and this motivated the development of our quasi-dense approach. Though the quasi-dense 3D points are the best we can achieve in 3D in terms of computability, but in image space, often the object image boundaries are expected for many geometric reconstruction. This means that the dense image segmentation could have been incorporated into the probabilistic framework. This can be done by extending the definition of labeling to all pixels. Let $L = L_X \cup L_I$ be the collective labeling for the joint points and all image pixels. Then the segmentation is given by

$$L = \arg\max_L P(L \mid X, V, I).$$

The probability $P(L \mid X, V, I)$ is factored into two terms

$$P(L \mid X, V, I) = P(L_I \mid X, V, I, L_X) P(L_X \mid X, V, I),$$

where the second term is the labeling model of joint points L_X, and the first term is image labeling L_I given the labeling of the joint points L_X or the propagation of L_X into image segmentations L_I. This propagation model can roughly be interpreted as a data augmentation problem (Tanner and Wong, 1987) in which L_X is treated as hidden variables. Usually an iterative algorithm could be adopted to this optimization problem, however, from the perspective of labeling and the fact that the visibility V of X is given, it is unnecessary to come back to re-estimate L_X from an estimate of L_I. This leads to the following segmentation propagation.

Given the labeling of the joint points L_X, the corresponding (quasi-dense) pixels in each view have been assigned the corresponding labels since the correspondence information and the visibility are provided. The probabilistic dependency relation of joint points and image segmentation is represented as a Bayesian network illustrated in Fig. 7(a). According to the conditional independence and Bayesian network properties, the Bayesian network can be exactly divided into n independent networks as shown in Fig. 7(b), where $\gamma_v = \{\mathbf{x}\} \subset X$ is the set of the joint points on which the labels of image I_v are dependent. This means that labeling other pixels in each view is dependent on those fixed labels and the image itself. It should be noted that the division holds only if the labels of joint points are given. Formally, $P(L_I \mid X, V, I, L_X)$ can be factorized into:

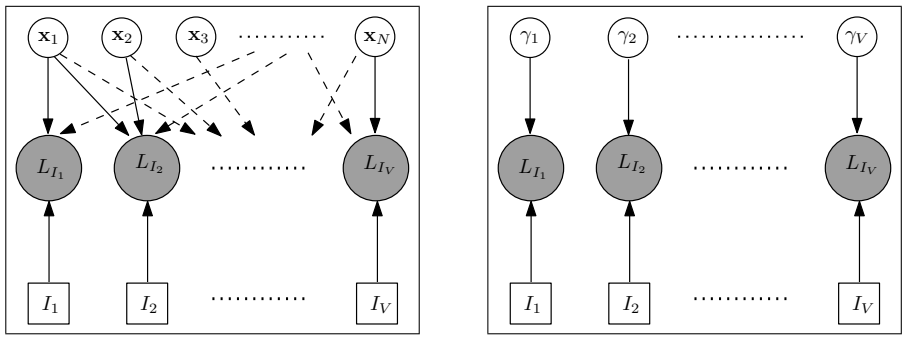
$$\begin{aligned}(L_{I_1}, \ldots, L_{I_n}) &= \arg\max P(L_{I_1}, \ldots, L_{I_n} \mid X, V, I, L_X) \\ &= \arg\max \prod_{v=1}^{n} P(L_{I_v} \mid X, V, I, L_X) \\ &= \arg\max \prod_{v=1}^{n} P(L_{I_v} \mid L_{\gamma_v}, I_v). \end{aligned} \quad (10)$$

The independence of views implies that it amounts to solving $L_I = \arg\max P(L_I \mid L_\gamma, I)$ if we drop the subscript v for simplicity. Again by adopting CRF to model the probabilistic segmentation model, we have

$$\begin{aligned}P(L_I \mid L_\gamma, I) \propto \exp -\{&E^c(L_I; L_\gamma) + \rho E^l(L_I; I) \\ &+ \zeta E^s(L_I; I)\},\end{aligned}$$

where $E^c(L_I; L_\gamma)$ is used to penalize the inconsistency between the labels of joint points and its corresponding image pixels, $E^l(L_I; I)$, called image compatibility term, is used to penalize the incompatibility between the labels and color information of pixels, and $E^s(L_I; I)$, called regularization term, is used to bias the smoothness of the labels of neighboring pixels.

This leads to a formulation that is very close to the image segmentation methods proposed in Boykov and Jolly (2001), Li et al. (2004), Blake et al. (2004) and Rother



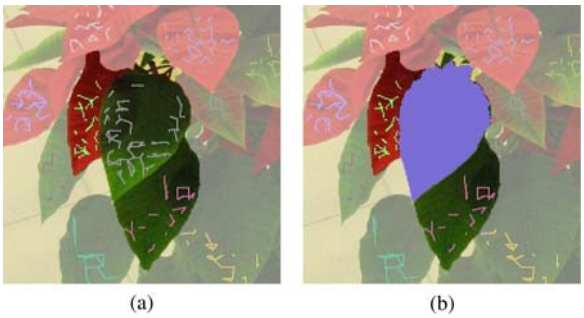
(a) (b)

Figure 7. (a) The network shows the conditional relation of the joint segmentation probability model for all views. The $a \rightarrow b$ means that b is conditionally dependent on a. (b) The network shows the independence factorization of the network in (a) according to the probability property.

et al. (2004). The difference is that the condition L_μ or the energy E^c term is provided differently in different methods. The existing methods put the user into the loop to provide an iterative and interactive condition by marking up the image, while our method uses the condition L_μ available from the given quasi-dense points and previously computed joint segmentation. We briefly specify the three energy terms to complete the description of the whole procedure.

Consistency constraint. Consistency term aims at keeping the consistency between the labels of image pixels and the associated joint points. It is given by $E^c(L_I; L_\gamma) = -\eta \sum_{i \in \gamma} \delta(l_k = \hat{l}_i)$, where k is the pixel index of the joint point i, and $\hat{l}_i$ is the label of the joint point. We impose a strict consistency constraint by setting $\eta = \infty$. It should be noted that the seeds L_γ are semi-dense as shown in Fig. 8(a), which makes labeling of the remaining pixels much more robust and effective.

Image compatibility. Image compatibility term is defined as $E^l(L_I; I) = -\sum_k \log p(c_k \mid l_k)$, where $p(c \mid l)$ is a GMM probabilistic model that describes the color distribution of the pixels with the label l. It is used to penalize the incompatibility between the label and color information of each pixel.



(a) (b)

Figure 8. (a) The superimposition of the projections of segmented 3D groups with one image. (b) The image segmentation of one group in blue, representing one leaf, from the associated group of 3D points.

Regularization. A regularization term is necessary to make the labels of neighboring pixels as smooth as possible. We take a Potts spatial energy model, $E^s(L_I; I) = -\sum_{(m,n)\in\mathcal{C}} a(m,n)\delta(l_m = l_n)$, where $\mathcal{C} = \{(m,n);\mid (x_m, y_m) - (x_n, y_n) \mid \leq d\}$ with d being the neighborhood size and typically set as 1, and the affinity between pixels $a(m,n) = \frac{1}{1+\epsilon}(\epsilon + \exp(-\frac{||c_m - c_n||^2}{2\sigma^2}))$, where ϵ is a dilution parameter for color contrast and is set as 0.1 by default, and σ is the standard deviation and estimated as $\sigma = \sqrt{\mathrm{E}(\mid c_m - c_n \mid^2)}$. This model avoids the natural tendency for segmentation boundaries to align with colors of high image contrast.

Computation. The overall objective function for segmentation propagation is given by

$$\begin{aligned} E &= E^c(L_I; L_\gamma) + \rho E^l(L_I; I) + \zeta E^s(L_I; I) \\ &= -\eta \sum_{i \in \gamma} \delta(l_k = \hat{l}_i) - \rho \sum_k \log p(c_k \mid l_k) \\ &\quad - \zeta \sum_{(m,n)\in\mathcal{C}} \delta(l_m = l_n) a(m,n). \end{aligned} \tag{11}$$

It is typically solved by graph cut (Boykov and Jolly, 2001). For the objects having apparent color similarity, for example when we would segment out individual leaves, the image compatibility term is not discriminative and can be ignored. The energy then reduces to

$$E = -\eta \sum_{i \in \gamma} \delta(l_k = \hat{l}_i) - \zeta \sum_{(m,n)\in\mathcal{C}} \delta(l_m = l_n) a(m,n).$$

This can still be optimized using graph cut. Different from existing methods in Boykov and Jolly (2001), Li et al. (2004), Rother et al. (2004), Kolmogorov et al. (2005), Sun et al. (2006) and Criminisi et al. (2006), here

only hard constraint, i.e. the consistency term, is involved. It appears to be insufficient for pixel labeling, but the number of the labeled pixels from the joint segmentation is larger than those in Boykov and Jolly (2001), Li et al. (2004), Rother et al. (2004) and Blake et al. (2004). The propagation from L_γ to the whole image is efficient using a graph cut algorithm as shown in Fig. 8(b).

5. Implementation and Results

Data acquisition. We typically capture about 35 images by moving around a foreground object as discussed in Lhuillier and Quan (2005). We choose two typical indoor scenes and three scenes of plants. The choice of plant images is motivated by the fact that the plants are omnipresent in many scenes and are reputed to be difficult in segmentation and modeling, and that the segmentation into individual leaves from pure images is not yet possible to the best of our knowledge even with intensive learning. On the other hand, if we want to obtain a realistic model of a plant as demonstrated in Quan et al. (2006), it is inevitable that a proper segmentation should be obtained. Obviously, the examples we choose have relatively large sizes of leaves for the given resolution of the cameras. For the dense foliage of the trees with small leaves, it exhibits texture-like properties and a different methodology should be developed.

The two indoor scenes are captured with 48 and 20 images of resolutions 972×1296 and 1024×768. The plants are captured with 35, 35 and 40 for nephthytis, poinsettia and schefflera. The image resolution is 1944×2592 (except for the poinsettia, which is 1200×1600). For the efficiency of structure from motion, we downsampled the images to 583×777 (for the poinsettia, to 600×800). It took approximately 10 mins for about 40 images on a 1.9GHz P4 PC with 1 GB of RAM. On average, we reconstructed about 100 thousands 3D points for the scene.

Outline of the algorithms. For each data set, the computation of the joint segmentation for 3D points and 2D images proceeds as follows:

1. Train the color distributions for foreground and background. We selected several views randomly. And for each selected view the depths of the pixels that have 3D corresponding points are calculated, and normalized to [0, 1]. Then two thresholds $d_f = \frac{1}{5}$ and $d_b = \frac{3}{5}$ are chosen to set the pixel whose depth is smaller than d_f as foreground pixel, and larger than d_b as background.
2. Optimize Eq. (6) using graph cut to obtain foreground and background separation. We set $\frac{\lambda}{\rho_1} = 2.5$ by default, the graph cut implementation in Boykov et al. (2001) is used.
3. Compute the second smallest eigenvector for the normalized Laplacian matrix $\mathbf{L} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{D} - \mathbf{W})\mathbf{D}^{-\frac{1}{2}}$ using sparse singular value decomposition (SVD) algorithm.
4. Threshold the eigenvector to produce a bi-partitioning of the 3D points. We chose 25 different values uniformly spaced within the range of the eigenvector as possible thresholds. Then we chose the one corresponding to a partition which minimizes the normalized cut value. The corresponding partition is accepted.
5. Recursively repeat step 3 and 4 for each partition until the normalized cut value is larger than 0.06.
6. Perform MPPCA to refine the 3D segments for the foreground points.
7. Segment each image by optimizing Eq. (11) using graph cut algorithm in Boykov et al. (2001) given segmented 3D points. And we fix $\eta = \infty$, $\rho = 1$ and $\zeta = 2.5$.
8. Fitting a specific geometric representation to each group, more details are given in the following paragraph.

Segmentation-based modeling. Each segmented group should be further processed to build an appropriate geometric representation as an object.

- If a group is representing a regular geometric object, for instances, a plane, a polyhedron, or a cylinder, it is straightforward to use some standard methods of fitting these well-defined geometric models to the given data.
- If a group is representing a smooth surface like a human head or a compact object, we could use a level set approach that integrates all joint points, image information and the object boundaries to build an implicit surface model (Lhuillier and Quan, 2005). Alternatively, we used a graph-cut approach that builds a surface model with more details but higher computational cost (Zeng et al., to appear).
- If a group is representing the specific hair of a given person, a combination of synthesis and analysis method could be used to reconstruct each hair fiber as a curve represented as a set of connected line segments by following the edge orientation in the images (Wei et al., 2005).
- If a group is representing an individual leaf of a plant, then we can build a generic leaf model for each plant, and we have developed a method of fitting a generic deformable model to the data in Quan et al. (2006).

We did not develop any specific fitting method in this paper, we used the combination of the above mentioned methods to build the final models illustrated in Figs. 2 and 13.

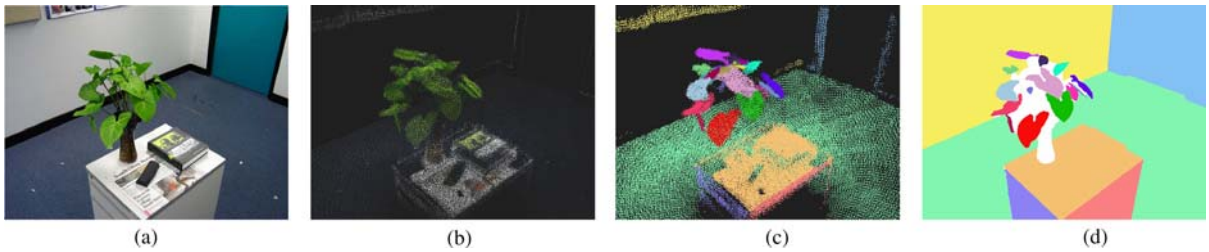


Figure 9. (a) The 3D points. (b) The joint segmentation results. (c) One of the original images. (d) One of the image segmentations by propagation.

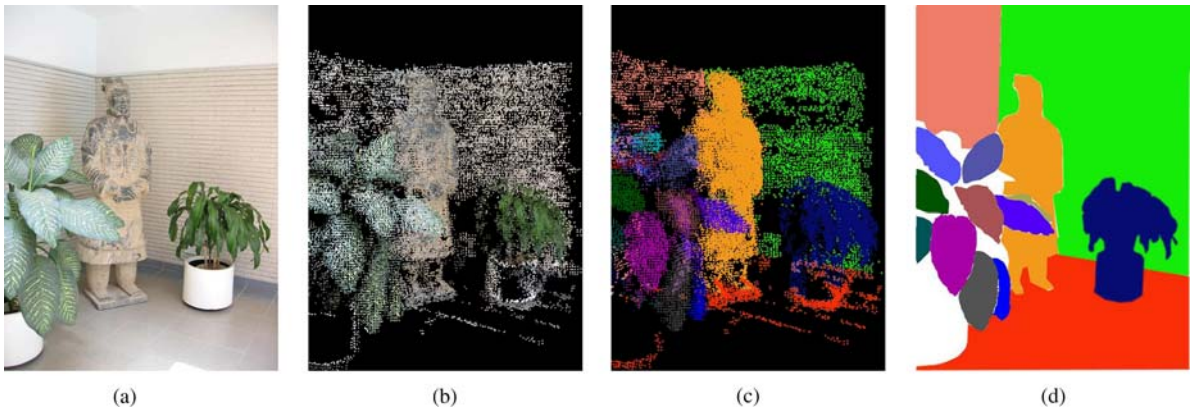


Figure 10. (a) The 3D points. (b) The segmented 3D groups by 3D segmentation. (c) One of the original images. (d) One of the image segmentations by propagation.

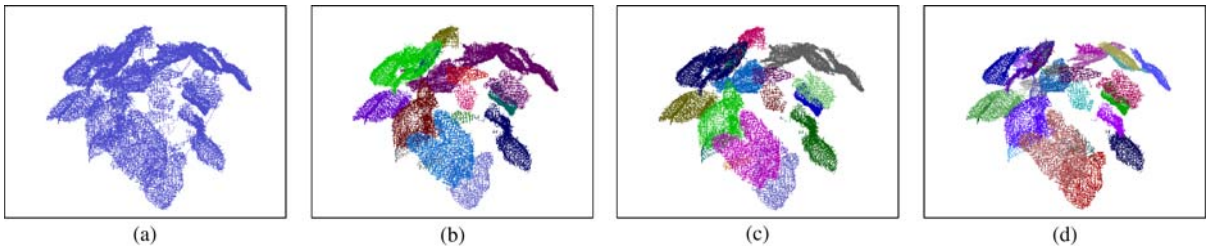


Figure 11. Two intermediate 3D segmentation results of the nephthytis example to illustrate the successive splitting of larger groups into the smaller ones.

Results. For the example scenario shown in Fig. 2, the joint segmentation results are given in Fig. 9. The walls, the small desk, and individual leaves have been successfully segmented, whereas the book and the eyeglasses box are not due to lack of discriminant reconstructed 3D points associated with them. These automatic results are then edited using a semi-automatic system as described in Quan et al. (2006) to create the final model of the scene rendered by Maya in Fig. 13.

Another example of scene shown in Fig. 10 segments out the different walls, the statue, and the leaves of one plant. The second smaller plant is segmented out as a whole not into leaves and flower pot, this is due to that the same scale, therefore the same parameters, is used for the whole scene, and it will be possible by changing the setting of the parameters.

For the difficult segmentation of a plant into leaves, when the leaves of the plant tend to be sufficiently large, that is the case of the nephthytis as shown in the first row of Fig. 12, the segmentation is relatively easier as the spatial distance is larger between the points of different leaves, and the image contour is strong as well. The segmentation results are excellent, since all visible leaves have been successfully partitioned. The intermediate results of the recursive strategy in step 3–4 of the above procedure are shown in Fig. 11. When the leaves of the plants become smaller such as in the cases of the poinsettia and schefflera shown in Fig. 12, the occlusion between leaves makes the segmentation more difficult, so the success rate is lower than the nephthytis example. The statistics of the results are reported in Table 1.

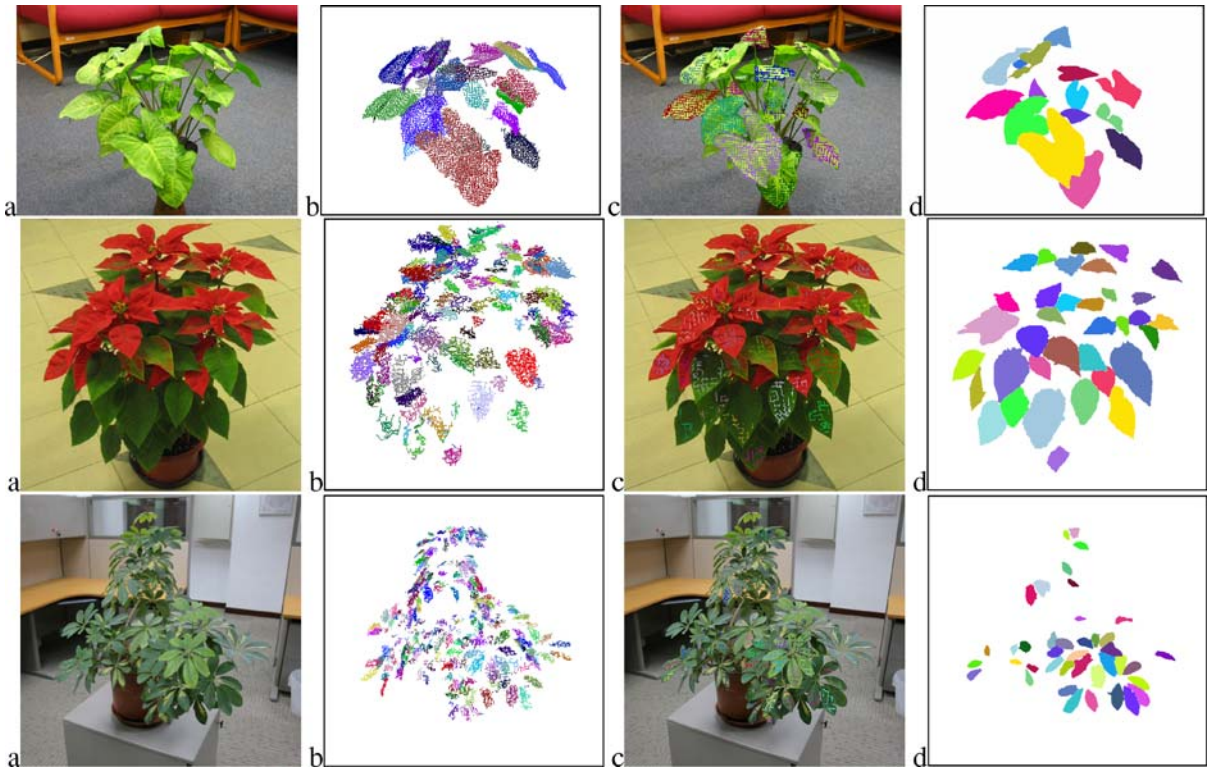


Figure 12. The first row is the nephthytis plant, the second is the poinsettia plant, and the third is the schefflera plant. (a) One of the original images. (b) Joint segmentation shown for 3D points, each group is coded with a different color. (c) Joint segmentation shown for the visible 2D points at one view. (d) Propagated image segmentation.

Figure 13. One of the original images of the schefflera plant on the left. Rendered at a similar viewpoint of the reconstructed full model of the plant on the right.

Table 1. The statistics for the 3D segmentation results on the three plants jointly using 3D and 2D information. The last row gives a rough estimate of the true number of the leaves for each plant by inspection.

	Nephthytis	Poinsettia	Schefflera
# images	35	35	40
# total 3D pts	128,000	103,000	118,000
# foreground pts	53,000	83,000	43,000
# segmented groups of the foreground	29	97	343
ground truth of # leaves	30	≈120	≈450

The segmentation results are not yet as perfect as we may expect, but fortunately many small objects such as individual leaves for the various plants have been successfully segmented out. The segmentation results have been successfully explored in a semi-automatic modeling approach in Quan et al. (2006) to produce realistic 3D models of the plants as shown in Fig. 13. Of course, the joint segmentation approach is general, not restricted to plants, and geometric fitting after the segmentation might be different for different types of objects.

6. Conclusion

By discussing the two concepts developed in the group led by Kanade, feature tracking for correspondences and factorization for reconstruction, we suggested that the quasi-dense approach overcomes the shortcomings of the existing approaches. Given the availability of both 3D point data and 2D image data, we proposed a joint segmentation approach that is formulated within a probabilistic framework. We proposed the approximate solution to the joint segmentation. The results obtained on real data could be explored in an interactive modeling approach as presented in Quan et al. (2006) for modeling purposes. The future directions include the development of more efficient computation methods of the proposed joint segmentation formulation.

Acknowledgments

The work was supported by Hong Kong RGC Grant HKUST6190/05E, NSFC/RGC Grant N-HKUST602/05E, and RGC Grant HKUST619006.

References

Blake, A., Rother, C., Brown, M., Pérez, P., and Torr, P.H.S. 2004. Interactive image segmentation using an adaptive GMMRF model. In *ECCV (1)*, pp. 428–441.

Boykov, Y. and Jolly, M. 2001. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *ICCV*, pp. 105–112.

Boykov, Y., Veksler, O., and Zabih, R. 2001. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239.

Criminisi, A., Cross, G., Blake, A., and Kolmogorov, V. 2006. Bilayer segmentation of live video. In *CVPR*.

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B:1–38.

Faugeras, O. 1992. What can be seen in three dimensions with an uncalibrated stereo rig? In Sandini, G. (Ed.), In *Proceedings of the 2nd European Conference on Computer Vision, Santa Margherita Ligure, Italy*, pp. 563–578. Springer-Verlag

Faugeras, O., Luong, Q., and Papadopoulo, T. 2001. *The geometry of multiple images*. The MIT Press, Cambridge, MA, USA.

Förstner, W. 1994. A framework for low level feature extraction. In *Proceedings of the 3rd European Conference on Computer Vision, Stockholm, Sweden*, pp. 383–394.

Fua, P. 1991. Combining stereo and monocular information to compute dense depth maps that preserve discontinuities. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence, Sydney, Australia*.

Gargallo, P. and Sturm, P. 2005. Bayesian 3D Modeling from Images Using Multiple Depth Maps. In *CVPR (2)*, pp. 885–891,

Harris, C. and Stephens, M. 1988. A combined corner and edge detector. In *Alvey Vision Conference*, pp. 147–151.

Hartley, R.I. 1992. Estimation of relative camera positions for uncalibrated cameras. In Sandini, G. (Ed.), In *Proceedings of the 2nd European Conference on Computer Vision, Santa Margherita Ligure, Italy*, pp. 579–587, Springer-Verlag.

Hartley, R.I. and Zisserman, A. 2000. *Multiple View Geometry in Computer Vision*. Cambridge University Press.

Urban, M., Matas, J., Chum, O., and Pajdla, T. 2002. Robust wide baseline stereo from maximally stable extremal regions. In *British Machine Vision Conference*, pp. 384–393.

Koenderink, J.J. and van Doorn, A.J. 1989. Affine structure from motion. Technical report, Utrecht University, Utrecht, The Netherlands, also appeared in *Journal of the Optical Society of America A*, 8(2):377–385, 1991.

Kolmogorov, V., Criminisi, A., Blake, A., Cross, G., and Rother, C. 2005. Bi-Layer segmentation of binocular stereo video. In *CVPR (2)*, pp. 407–414.

Kolmogorov, V. and Zabih, R. 2002. Multi-camera scene reconstruction via graph cuts. In *ECCV (3)*, pp. 82–96.

Lafferty, J.D., McCallum, A., and Pereira, F.C.N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pp. 282–289.

Laveau, S. 1996. *Géométrie d'un système de N caméras. Théorie, estimation, et applications*. Thèse de doctorat, École Polytechnique.

Lhuillier, M. 1998. Efficient dense matching for textured scenes using region growing. In *Proceedings of the ninth British Machine Vision Conference, Southampton, England*, pp. 700–709.

Lhuillier, M. and Quan, L. 2002. Image-based rendering by match propagation and joint view triangualtion. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1140–1146.

Lhuillier, M. and Quan, L. 2005. A quasi-dense approach to surface reconstruction from uncalibrated images. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 27(3):418–433.

Li, Y., Sun, J., Tang, C., and Shum, H. 2004. Lazy snapping. In *Proceedings of ACM SIGGRAPH.*, pp. 303–308.

Lowe, D. 2004. Distinctive image feature from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Lucas, B.D. and Kanade, T. 1981. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*.

Malik, J., Belongie, S., Leung, T.K., and Shi, J. 2001. Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 43(1):7–27.

Mohr, R., Quan, L., and Veillon, F. 1995. Relative 3D reconstruction using multiple uncalibrated images. *International Journal of Robotic Research*, 14(6):619–632.

Mohr, R., Quan, L., Veillon, F., and Boufama, B. 1992. Relative 3D reconstruction using multiple uncalibrated images. Technical Report RT 84-I-IMAG LIFIA 12, LIFIA–IRIMAG.

Moravec, H. 1979. Visual mapping by a robot rover. In *Proceedings of the 6th International Joint Conference on Artificial Intelligence, Tokyo, Japan*, pp. 598–600.

Moravec, H. 1981. Obstable avoidance and navigation in the real world by a seeing robot rover. Technical report CMU-RI-tr-3, Carnegie Mellon University.

Nister, D. 2001. *Automatic Dense Reconstruction from Uncalibrated Video Sequences*. Ph.d. thesis, NADA, KTH, Sweden.

Patras, I., Hendriks, E.A., and Lagendijk, R.L. 2001. Video segmentation by MAP labeling of watershed segments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(3):326–332.

Pollard, S.B., Mayhew, J.E.W., and Frisby, J.P. 1985. PMF: a stereo correspondence algorithm using a disparity gradient limit. *Perception.* 14, pp. 449–470.

Pollefeys, M., Koch, R., and Van Gool, L. 1998. Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters. In *Proceedings of the 6th International Conference on Computer Vision, Bombay, India*, pp. 90–95.

Quan, L. 1995. Invariants of six points and projective reconstruction from three uncalibrated images. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 17(1):34–46.

Quan, L., Tan, P., Zeng, G., Yuan, L., Wang, J., and Kang, S.B. 2006. Image-based plant modeling. In *Proceedings of ACM SIGGRAPH*.

Rother, C., Kolmogorov, V., and Blake, A. 2004. GrabCut: interactive foreground extraction using iterated graph cuts. In *Proceedings of ACM SIGGRAPH.*, pp. 309–314.

Shi, J. and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905.

Shi, J. and Tomasi, C. 1994. Good features to track. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Seattle, Washington, USA*, pp. 593–600.

Sturm, P. and Triggs, B. 1996. A factorization based algorithm for multi-image projective structure and motion. In B. Buxton and R. Cipolla, editors, *Proceedings of the 4th European Conference on Computer Vision, Cambridge, England*, volume 1065 of *Lecture Notes in Computer Science*, pp. 709–720. Springer-Verlag.

Sun, J., Zhang, W., Tang, X., and Shum, H. 2006. Background Cut. In *ECCV*.

Tanner, M.A. and Wong, W.H. 1987. The calculation of posterior distributions by data augmentation (with discussion). In *Journal of the American Statistical Association*, 82, 528–550.

Tipping, M.E. and Bishop, C.M. 1999. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482.

Tomasi, C. 1991. *Shape and Motion from Image Streams: a Factorization Method*. PhD thesis, Carnegie Mellon University, USA.

Tomasi, C. and Kanade, T. 1991. Detection and tracking of point features. Technical report CMU-CS-91-132, Carnegie Mellon University.

Tomasi, C. and Kanade, T. 1992. Shape and motion from image streams under orthography: A factorization method. *International Journal of Computer Vision*, 9(2):137–154.

Torr, P.H.S., Szeliski, R., and Anandan, P. 2001. An integrated bayesian approach to layer extraction from image sequences. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(3):297–303.

Triggs, B. 1996. Factorization methods for projective structure and motion. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, San Francisco, California, USA*, pp. 845–851.

Triggs, B. 1997. Autocalibration and the absolute quadric. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Puerto Rico, USA*, pp. 609–614. IEEE Computer Society Press.

Triggs, B. 2004. Detecting keypoints wiht stable position, orientation and scale under illumination changes. In *European Conference on Computer Vision*. Springer-Verlag.

Triggs, B., McLauchlan, P.F., Hartley, R.I., and Fitzgibbon, A. 2000. Bundle ajustment—a modern synthesis. In Triggs, B., Zisserman, A., and Szeliski, R. (Eds.), *Vision Algorithms: Theory and Practice*, volume 1883 of *Lecture Notes in Computer Science*, pp. 298–372. Springer-Verlag.

Tu, Z. and Zhu, S.C. 2002. Image segmentation by data-driven Markov chain Monte Carlo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):657–673.

Tuytelaars, T. and Van Gool, 2000. Wide baseline stereo based on local, affinely invariant regions. In *British Machine Vision Conference*, pp. 412–422.

Wang, J.Y.A., and Adelson, E.H. 1994. Representing moving images with layers. *IEEE Transactions on Image Processing*, 3(5):625–638.

Wei, Y., Ofek, E., Quan, L., and Shum, H. 2005. Modeling hair from multiple views. *ACM Transactions on Graphics* (TOG), Proceedings of ACM SIGGRAPH 2005 (SIGGRAPH), vol. 27, no. 3.

Wills, J., Agarwal, S., and Belongie, S. 2003. What went where. In *CVPR (1)*, pp. 37–44,

Xiao, J. and Shah, M. 2004. Motion layer extraction in the presence of occlusion using graph cut. In *CVPR (2)*, pp. 972–979.

Zabih, R. and Kolmogorov, V. 2004. Spatially coherent clustering using graph cuts. In *CVPR (2)*, pp. 437–444.

Zeng, G., Paris, S., Quan, L., and Sillion, F. to appear. Accurate and scalable surface representation and reconstruction from images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *(IEEE TPAMI)*.

Zhang, Z., Deriche, R., Faugeras, O.D., and Luong, Q.T. 1995. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78(1–2):87–120. Appeared in October 1995, also INRIA Research Report No.2273, May 1994.

Zhu, X. and Lafferty, J. 2005. Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. In *ICML*, pp. 1052–1059.