Probabilistic 3D Occupancy Flow with Latent Silhouette Cues

Li Guan, Jean-Sébastien Franco, Edmond Boyer, Marc Pollefeys

Probabilistic 3D Occupancy Flow with Latent Silhouette Cues

Li Guan
UNC-Chapel Hill, U.S.A.
lguan@cs.unc.edu

Jean-Sébastien Franco
LaBRI - INRIA Sud-Ouest
University of Bordeaux, France
jean-sebastien.franco@labri.fr

Edmond Boyer
Universities - INRIA Grenoble, France
edmond.boyer@inrialpes.fr

Marc Pollefeys
ETH-Zürich, Switzerland
UNC-Chapel Hill, U.S.A.
marc.pollefeys@inf.ethz.ch

Abstract

In this paper we investigate shape and motion retrieval in the context of multi-camera systems. We propose a new low-level analysis based on latent silhouette cues, particularly suited for low-texture and outdoor datasets. Our analysis does not rely on explicit surface representations, instead using an EM framework to simultaneously update a set of volumetric voxel occupancy probabilities and retrieve a best estimate of the dense 3D motion field from the last consecutively observed multi-view frame set. As the framework uses only latent, probabilistic silhouette information, the method yields a promising 3D scene analysis method robust to many sources of noise and arbitrary scene objects. It can be used as input for higher level shape modeling and structural inference tasks. We validate the approach and demonstrate its practical use for shape and motion analysis experimentally.

1. Introduction

Building 4D space-time representations of scenes observed from multiple calibrated views is a major challenge in computer vision, for acquisition of full 3D sequences from images. They are relevant to many fields in research and industry, for free-viewpoint video, automatic 3D shape and human performance acquisition methods, virtual reality and HCI applications, 3D shape matching and recognition. Efficient representations are needed, to track and build time-coherent 3D shape geometry, analyze and acquire 3D motion of subjects in the scene. Often such problems are cast as an estimation of a 3D surface, wether tracked using a 3D template[27, 8], or temporally aligned [24]. Low level approaches exist to estimate 3D motion and help surface estimation and tracking, such as scene flow techniques [25] to estimate local surface displacements, but they generally rely on specific surface geometry or appearance assumptions for scene objects. All such approaches can run into limitations due to apparent topology changes and self-occlusion of observed surfaces, inability to deal with arbitrary scenes or objects, image noise and variability in outdoor, uncontrolled acquisition environments. Probabilistic occupancy grids are an interesting representation to address many of these problems because they assume little about the observed scene and can be associated to generative models of images to

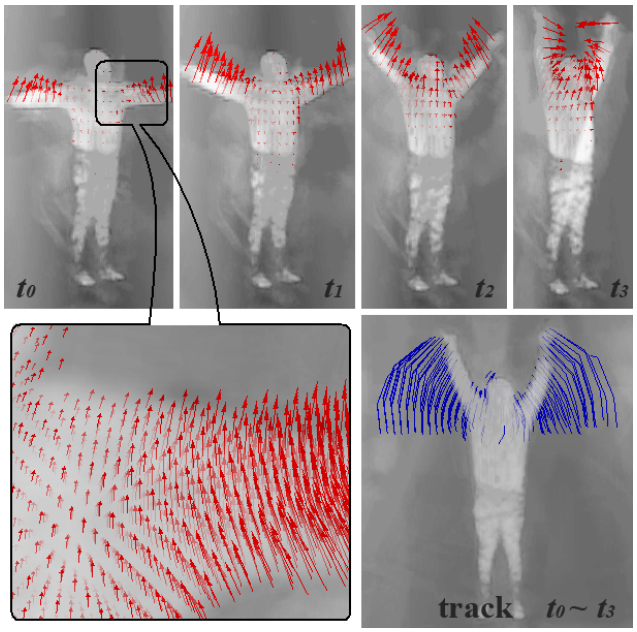

Figure 1. Recovered dense flow and the 3D probabilistic occupancy grid. The volumes are rendered with near-transparent α-value to voxels of lower than 0.98 occupancy probability. The motion vectors are in red. Part of t_0's motion field is magnified at the bottom left. The cumulated point tracks over $t_0 - t_3$ in blue as on the bottom right. **Best viewed in color.**

explicitely deal with noise. They have been shown to be particularly useful not only in computer vision [3, 1, 11], but also in other fields such as robotics [10]. The ability to propagate such probabilistic representations over time is required when dealing with temporal sequences, but is a challenging scientific problem seldom addressed by existing methods, which usually address updates of explicit shapes or surfaces.

In this paper we propose an optimal (MAP) and efficient solution to address occupancy grid updates. We apply this analysis to the case of silhouette cues, but other types of inputs could be used in the framework. To solve this problem we model the intuition that shape and motion estimation are mutually helpful, and thus naturally solve the resulting probabilistic problem using an iterative, EM algorithm. This enables the method to cooperatively estimate a probabilistic 3D shape representation and dense 3D scene flow for a pair of consecutive frame-sets in the sequence. As such, the proposed framework explores what minimal constraints and data can be used for 3D motion analysis from multiple views in difficult conditions, and provides a new tool which can be used for higher level analysis such as kinematic chain discovery, shape tracking and incremental improvement across time.

1.1. Related works

Image-based modeling The problem of 3D shape acquisition across time was first approached in a purely frame-by-frame manner. Photocoherence [3], and wide-baseline stereo [21] are widely useful to reconstruct surface elements, but have intrinsic limitations and computational overhead related to object self-occlusion, degraded color inputs and often simplistic BRDF assumptions. The overhead of these methods pays off for well-textured regions but yields no additional information with poorly textured data, often the case with casual clothing (e.g. Fig. 1 and Fig. 3(2)&(3)). Silhouette-based methods [18, 19] have complemented such techniques as they are generally fast and robust under various types of image noise, and unconvenient appearance properties, such as weak texture. Most aforementioned shape modeling approaches focus on surface representations, yet alternative representations, such as volumetric probability grids, have emerged to improve robustness to noise of various methods including photometric space carving [3], by removing some premature hard decisions on shape location. The use of such representations with latent silhouette data has been shown to be particularly robust for difficult, outdoor environments [11, 13, 14, 1], a property we wish to leverage in this paper. But, to our knowledge, no paper addresses temporal updates of probabilistic occupancies from images as proposed here.

Shape tracking approaches. Recently, mesh tracking methods have proven to be successful for time-consistent shape acquisition and refinement. Many such methods fit existing, fixed-topology mesh model templates [8] to image data. These methods are however often particularized for the case of a specific shape, usually human [27], by underlying geometric or kinematic assumptions. Some methods do aim for more general surfaces and can sometimes deal with surface topology changes [24]. To constrain surface estimation, the methods use a variety of image cues, such as dense optical flow [8], sparse feature matches [24]. Alternative methods exist, e.g. Cheung *et al.* combine voxel-based representations with silhouette inputs, albeit in the case of rigid or articulated objects [6]. Because of the inherent difficulties and ill-posed nature of shape tracking, most methods currently use some form of manual input or specific initialization.

Remarkably, a large majority of these methods use silhouette-based constraints to stabilize estimation, given that real-life scenes generally provide too few reliable sparse matches to constrain surface estimation, as noted in [7]. In particular, some surface-tracking methods actually use silhouette data alone [27], a testament to their constraining power, which we wish to leverage in our method. Also, the vast majority of 4D shape tracking methods has only been tested in completely controlled environments, where silhouettes and features are easy to obtain. This suggests the large difficulty in using them with uncontrolled inputs and outdoor acquisitions. These methods could thus potentially benefit from our probabilistic analysis, which can produce results outdoors with no geometric assumptions about the scene, to constrain or initialize mesh motion estimates.

Scene flow approaches The method we propose is related to scene flow approaches. These methods compute motion fields associated to various shape representations including voxels [25], level sets [20], stereo disparity maps [28], surfels [5], or meshes [8]. Scene flow approaches require addressing self-occlusion as they estimate surface properties. This is often dealt with by assuming an underlying surface representation is already computed [25, 8] or built simultaneously [20]. A large family of scene flow methods rely on the estimation of spatial derivatives of the image signal, and rely on restrictive BRDF assumptions are used for method simplicity. As noted in [22], flow-based approaches are generally limited to small displacements, as a consequence of finite difference approximations of derivatives. 6D carving of spatiotemporal voxel pairs based on photo-consistence and bounded voxel motion assumptions is an alternative [26], but leads to combinatorial searching. Recent methods such as [20] address BRDF and derivative approximation related issues, but require high computational complexity algorithms to estimate a surface model and still

strongly rely on high texture content to produce good results. Our flow analysis solves a similar yet new problem, with two key differences and contributions: (1) solving the flow problem in a 3D volume rather than a 2D manifold to improve robustness, genericity and deal with occupancy grid updates; (2) use of different cues obtained from silhouettes exclusively, while showing that usable information can still be obtained, including in the case of degraded inputs and difficult, low-textured subjects.

1.2. Overview

We cast the problem as the simultaneous registration and probability update of two time-consecutive occupancy grids. Our formulation and assumptions (§2) are analog to probabilistic interpretations of 2D optical methods [23] extended to 3D volumes and silhouette inputs. In particular we only use the spatial continuity of the motion field as regularization. Initialization is automatic and no specific kinematic or structural assumptions are used. Shape and motion are jointly estimated with an Expectation Maximization algorithm (§3), which alternates between estimating voxel occupancy probabilities in the E-step (§3.1), and finding a motion field best estimate in the M-step (§3.2). Discretizing motion possibilities casts the M-step as a multi-label MRF, which we efficiently solve using a coarse-to-fine approach allowing large displacements (§4). The method is validated on several indoor, outdoor, and synthetic datasets (§5).

2. Problem Formulation

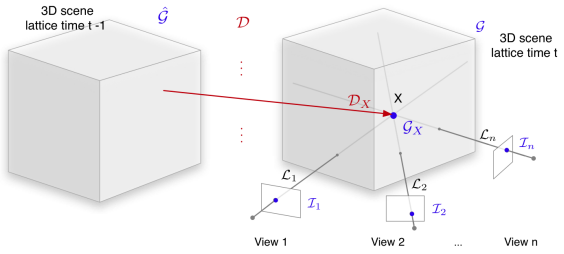

Figure 2. Overview of main statistical variables and geometry of the problem. $\mathcal{G}_X$ is the occupancy at voxel X.

We represent the scene with a 3D lattice of points in space (Fig. 2), denoted as $\mathcal{X}$. At time t, we observe a set of images $\mathcal{I}$, specifically $\mathcal{I}_1, \cdots, \mathcal{I}_n$ from n camera views with known projection matrices. We associate to each point $X \in \mathcal{X}$ a binary occupancy state, empty or occupied, denoted by $\mathcal{G}_X \in \{0, 1\}$. The conjunction of all grid states is noted $\mathcal{G}$. We shall use the previous grid state from time $t-1$ given a certain time discretization, noted $\hat{\mathcal{G}}$. The motion of matter from $t-1$ to t is represented by a displacement field $\mathcal{D}$. Specifically, we associate to each point X the vector $\mathcal{D}_X$ that displaces matter from location $X - \mathcal{D}_X$ to X between time $t-1$ and t. Since no surface representation is used, we assume the motion field is defined everywhere in space.

2.1. Expected properties

As the intent is to estimate the physical motion of objects underlying to the probability grid, continuity of the motion field should be encouraged for probably occupied regions, and is indifferent in probably empty regions. To be physically correct, the motion field should generally register the two object instances at $t-1$ and t. In practice this translates in our model to a propagation probability term $p(\mathcal{G}_X|\hat{\mathcal{G}}_{X-\mathcal{D}_X})$ (see §2.2) that favors mapping a probably occupied voxel at $t-1$ to a probably occupied voxel at t. This does not necessarily mean it should only map voxel pairs with exact probability *values*: occupancy probabilities are not equivalent to a physical opacity measure. Note that the estimated probability of voxels at t intuitively depends on the silhouette cues in images at t, but should also depend on probabilities of voxels mapped at $t-1$ to allow temporal filtering of occupancies. Thus a given choice of motion field influences the resulting occupancy probabilities, and vice-versa, naturally leading to an iterative EM formulation.

2.2. Joint Distribution

The relationship between the different factors of the system can be modeled by the joint probability of the variables: $p(\hat{\mathcal{G}}\mathcal{G}\mathcal{D}\mathcal{I})$, which we decompose in Eq. (1) with the following intuitions: occupancies $\mathcal{G}$ only depends on the displacements $\mathcal{D}$ and previous occupancies $\hat{\mathcal{G}}$. To predict images $\mathcal{I}$, only occupancies $\mathcal{G}$ at time t are needed. We more specifically assume *conditional* independence of voxel occupancies $\mathcal{G}_X$ *given the knowledge of* $\mathcal{D}_X$ and the previous occupancy before displacement, $\hat{\mathcal{G}}_{X-\mathcal{D}_X}$. This is analog to classic 2D optical flow formulations and their probabilistic interpretation given in [23], where pixel observations are predicted given only the optical flow vector at this pixel and the color of the previous frame's displaced pixel. Also, we assume that pixels to which a voxel center X projects have silhouette measurements that can be independently associated to that voxel. Thus no dependencies between voxels of the same viewing line need to be considered, which is completely analog to deterministic silhouette-based methods. Additionally assuming *conditional* independence of a voxel X's pixel measurements *given the knowledge of* the voxel's state $\mathcal{G}_X$, we can express $p(\hat{\mathcal{G}}\mathcal{G}\mathcal{D}\mathcal{I})$ as a product over the voxels, images and pixels:

$$p(\mathcal{D}) \prod_X \left(p(\hat{\mathcal{G}}_{X-\mathcal{D}_X}) p(\mathcal{G}_X|\hat{\mathcal{G}}_{X-\mathcal{D}_X}) \prod_i p(\mathcal{I}^i_{\mathsf{P}X}|\mathcal{G}_X) \right), \tag{1}$$

where i is the camera view index. $\mathcal{I}^i_{\mathsf{P}X}$ is the color of X's pixel projection in image i. For convenience, we later de-

note time t's measurement term $\mathbf{\Phi}(\mathcal{G}_X) = \prod_i p(\mathcal{I}^i_{\mathsf{p}X}|\mathcal{G}_X)$. The term $p(\mathcal{D})$ models the prior over the 3D motion field, used to regularize the field as described in §3.2. As we assume probabilistic inference information is already available for the previous time step $t-1$ for $\hat{\mathcal{G}}_X$, we treat $\hat{\mathcal{G}}_X$ as a latent variable of our problem. This enables to retain the probabilistic information $p(\hat{\mathcal{G}}_X)$ from $t-1$ by marginalizing out $\hat{\mathcal{G}}_X$ in all subsequent inferences.

3. Estimating 3D Motion and Occupancy

To solve the estimation problem we focus on estimating the Maximum A Posteriori (MAP) of $p(\mathcal{D}|\mathcal{I})$, treating $\mathcal{D}$ as our parameter set and $\mathcal{G}$ as our latent variable set. While EM was intially conceived for maximum likelihood problems, we need to use the MAP-EM generalization to incorporate a prior over $\mathcal{D}$, necessary to model motion field continuity. MAP-EM was shown to have identical convergence properties as EM [9]. The MAP-EM's goal is to find the optimal motion field d^* such that:

$$d^* = \text{argmax}_{\mathcal{D}}\ P(\mathcal{D}) \quad \text{with} \quad P(\mathcal{D}) = p(\mathcal{D}|\mathcal{I}). \tag{2}$$

This goal is to be achieved iteratively starting from an initial guess d^0, by building a sequence of motion field estimates $d^0, d^1, \cdots, d^*$ which increase the log-posterior objective function $P(\mathcal{D})$, i.e. $P(d^0) \leq \cdots \leq P(d^*)$ (M-Step). This is usually obtained at each step by maximizing a lower bound of $P(\mathcal{D})$, whose maximum coincides with an analytically simpler function $Q(\mathcal{D}|d^k)$ defined as follows [9]:

$$\begin{aligned} Q(\mathcal{D}|d^k) &= E_{\mathcal{G}|\mathcal{I},d^k}\{\ln p(\mathcal{I},\mathcal{G},\mathcal{D})\} \\ &= \sum_{\mathcal{G}} p(\mathcal{G}|\mathcal{I},d^k) \ln p(\mathcal{I},\mathcal{G},\mathcal{D}). \end{aligned} \tag{3}$$

The **E-step** first evaluates $p(\mathcal{G}|\mathcal{I},d^k)$ of Eq. (3), i.e. the grid occupancy probabilities given images $\mathcal{I}$ and the previously predicted displacement d^k. Then, the M-Step obtains the next motion field estimate d^{k+1} by maximizing $Q(\mathcal{D}|d^k)$ which garantees an increase of the log-posterior:

$$\textbf{M-Step:} \qquad d^{k+1} = \text{argmax}_{\mathcal{D}}\ Q(\mathcal{D}|d^k). \tag{4}$$

3.1. E-step: Occupancy Probability Update

In order to compute voxel probabilities at a given EM iteration, we need to express $p(\mathcal{G}|\mathcal{I},d^k)$ in terms of the joint probability distribution (1). We use Bayes' rule (5) and refactor the summations (6) to depending terms, with $\propto$ denoting proportionality up to a unit normalization factor:

$$p(\mathcal{G}|\mathcal{I},d^k) \propto \sum_{\hat{\mathcal{G}}} p(\hat{\mathcal{G}}\mathcal{G}d^k\mathcal{I}) \tag{5}$$

$$\propto \prod_X \mathbf{\Phi}(\mathcal{G}_X) \cdot \sum_{\hat{\mathcal{G}}_{X-d^k_X}} p(\hat{\mathcal{G}}_{X-d^k_X}) p(\mathcal{G}_X|\hat{\mathcal{G}}_{X-d^k_X}), \tag{6}$$

where $\sum_{\hat{\mathcal{G}}_{X-d^k_X}} p(\hat{\mathcal{G}}_{X-d^k_X}) p(\mathcal{G}_X|\hat{\mathcal{G}}_{X-d^k_X})$ sums possibilities over occupancy states of the voxel $X-d^k_X$ that has been mapped to X through displacement d^k_X. For simplicity we set $p(\mathcal{G}_X|\hat{\mathcal{G}}_{X-d^k_X})$ deterministically: if the previous voxel $X-d^k_X$ was occupied (resp. empty), then once displaced to X it is still occupied (resp. empty) with probability 1. Expression (6) then becomes:

$$p(\mathcal{G}|\mathcal{I},d^k) \propto \prod_X \left(\mathbf{\Phi}(\mathcal{G}_X) \cdot p([\hat{\mathcal{G}}_{X-d^k_X} = \mathcal{G}_X]) \right). \tag{7}$$

For the purpose of providing a probabilistic shape estimate, each voxel X's probability after an E-step can thus be identified as $p(\mathcal{G}_X|\mathcal{I},d^k) \propto \mathbf{\Phi}(\mathcal{G}_X) \cdot p([\hat{\mathcal{G}}_{X-d^k_X} = \mathcal{G}_X])$, the product of current observation terms at time t, with the probability of the voxel mapped to X from $t-1$.

$\mathbf{\Phi}(\mathcal{G}_X)$ can be computed by expliciting the image formation terms $p(\mathcal{I}^i_{\mathsf{p}X}|\mathcal{G}_X)$. For every pixel x in every image, we assume the parameters $\mathcal{B}$ of a background model have been learned offline from images of a quasi-static scene with no object of interest. Thus $p(\mathcal{I}^i_{\mathsf{p}X}|\mathcal{G}_X)$ can be set as follows:

$$p(\mathcal{I}^i_{\mathsf{p}X}|\mathcal{G}_X) = p(\mathcal{G}_X{=}0)p(\mathcal{I}^i_{\mathsf{p}X}|\mathcal{B}) + p(\mathcal{G}_X{=}1)\mathcal{U}(\mathcal{I}^i_{\mathsf{p}X}),$$

where $\mathcal{U}(\mathcal{I}^i_{\mathsf{p}X})$ is the uniform distribution over pixel color space, used to model the appearance of objects of interest since we use no information about it, and $p(\mathcal{I}^i_{\mathsf{p}X}|\mathcal{B})$ is the probability of $\mathcal{I}^i_{\mathsf{p}X}$ to be drawn from the background model $\mathcal{B}$. $p(\mathcal{I}^i_{\mathsf{p}X}|\mathcal{B})$ can be, for instance, a Normal or Gaussian Mixture Model distribution. The silhouette information is latent in this representation and does not require any binary segmentation decision.

3.2. M-step: 3D Motion Field Update

To optimize Eq. (4), we need to expand the expression of $Q(\mathcal{D}|d^k)$ in Eq. (3). The distribution $p(\mathcal{I},\mathcal{G},\mathcal{D})$ can be computed by marginalizing Eq. (1) over $\hat{\mathcal{G}}$, and simplified similarly to Eq. (6):

$$p(\mathcal{I},\mathcal{G},\mathcal{D}) \propto p(\mathcal{D}) \prod_X \mathbf{\Phi}(\mathcal{G}_X) \cdot p([\hat{\mathcal{G}}_{X-\mathcal{D}_X} = \mathcal{G}_X]). \tag{8}$$

Taking the logarithm of Eq. (8) to compute $Q(\mathcal{D}|d^k)$, and noting that the $\mathbf{\Phi}(\mathcal{G}_X)$ term doesn't depend on $\mathcal{D}$, the M-step becomes:

$$\begin{aligned} d^{k+1} = \text{argmax}_{\mathcal{D}}\ &\ln(p(\mathcal{D})) \\ &+ \sum_X \sum_{\mathcal{G}_X} p(\mathcal{G}_X|\mathcal{I},d^k) \cdot \ln p([\hat{\mathcal{G}}_{X-\mathcal{D}_X} = \mathcal{G}_X]), \end{aligned} \tag{9}$$

where $p(\mathcal{G}_X|\mathcal{I},d^k)$ is computed in the E-step (Eq. (7)).

To model 3D motion field continuity, we choose $p(\mathcal{D})$ to be a Markov Random Field. The M-step in our EM framework becomes a standard first-order MRF MAP problem.

From time t to $t+1$, if we *quantize* the displacement possibilities at every point to n displacement options denoted as a label set $\mathcal{L} = \{l^1, \cdots, l^n\}$, then we can rewrite Eq. (9) as a standard graph optimization with the following energy:

$$E_{MRF} = \sum_X \sum_{Y \in \mathcal{N}(X)} E_{XY}(l_X, l_Y) + \sum_X E_X(l_X), \quad (10)$$

where $\mathcal{N}(X)$ is the neighborhood system of point X in the 3D graph. In Eq. (10), $\sum_X \sum_{Y \in \mathcal{N}(X)} E_{XY}(l_X, l_Y)$ are the binary terms, and $\sum_X E_X(l_X)$ are the unary terms. They correspond to the *negative* of $\ln(p(\mathcal{D}))$ and $\sum_X \sum_{\mathcal{G}_X} p(\mathcal{G}_X|\mathcal{I}, d^k) \cdot \ln p([\hat{\mathcal{G}}_{X-\mathcal{D}_X} = \mathcal{G}_X])$ in Eq. (9) respectively.

The M-step in our EM framework becomes a discrete multi-labeling problem, with the goal of computing a labeling $L \in \mathcal{L}^{|\mathcal{X}|}$, which assigns each grid node $X \in \mathcal{X}$ a label from $\mathcal{L}$ such that the energy E_{MRF} is minimized. Thus

$$d^{k+1} = \text{argmin}_L E_{MRF}. \quad (11)$$

The solution to this MRF thus provides the updated displacement field in the EM iteration. We give further details on MRF implementation in the sections below.

4. Motion Field Optimization

Because of the large state space and ill-posed nature of the problem, and because MAP-EM has the potential to converge to unwanted local minima, additional steps must be taken to ensure convergence. We first review precisely how to regularize the motion field and ensure its continuity (§4.1). The resulting energy function can be optimized using Fast-PD approaches [17] (details in §4.2). We propose to apply Fast-PD in a coarse-to-fine approach for method stability, efficiency and convergence (§4.3).

4.1. Motion Field Properties

To ensure continuity of the motion field, we use the binary terms of the graph to enforce smoothness. We could define the pairwise energy function V_{XY} in Eq. (10) as a distance function computing the magnitude of vector differences [12]:

$$E_{XY}(l_X, l_Y) = \lambda_{XY} \left| \boldsymbol{d}(l_X) - \boldsymbol{d}(l_Y) \right|^{0.8}, \quad (12)$$

where λ_{XY} is a weighting factor, $\boldsymbol{d}(l)$ is the motion vector that label l represents and the index 0.8 is specially chosen to be less than one, motivated by statistics of velocity difference distribution studied for optical flow constraints [23].

However, as pointed out in [12], a more desirable pairwise energy term can be defined specifically to avoid overly fluid-like deformations in the case of iterative, coarse-to-fine approaches:

$$E_{XY}(l_X, l_Y) = \lambda_{XY} \left| \boldsymbol{D}_X + \boldsymbol{d}(l_X) - \boldsymbol{D}_Y - \boldsymbol{d}(l_Y) \right|^{0.8}, \quad (13)$$

where $\boldsymbol{D}_X$ and $\boldsymbol{D}_Y$ are the motions that have been recovered at location X and Y from previous iterations.

4.2. Fast-PD Optimization

The minimization of Eq. (10) in our M-step can be solved by discrete graph optimization schemes. We choose the Fast-PD approach [17], which builds upon principles drawn from the duality theory of linear programming in order to efficiently derive almost optimal solutions for a very wide class of NP-hard MRFs [16]. Indeed this approach has several advantages: it is faster than state-of-the-art graph cut α-expansion methods [2] and guarantees an optimality bound. In addition, it handles cost functions with arbitrary pair-wise potentials, lifting the *submodularity* constraint of previous approaches [15]. This gives us freedom to use the more elaborate forms of motion field local properties $E_{XY}(l_X, l_Y)$, such as the one we use in Eq. (13).

4.3. Coarse-to-Fine Approach

We opt for a coarse-to-fine approach and parametrization used for 3D volumetric registration in medical imaging [12]. We initialize the EM with a coarse global translation registration of grids. We then embed 3D space in a 3D B-Spline free form deformation (FFD) controlled by a uniform grid of sparse control points. At each chosen scale, the MRF previously identified is solved for the control points, and initialization for finer scales obtained by interpolation of the coarser scale. The control points are allowed discrete displacements possibilities $\mathcal{L}$ in the cubic range $[-d_s/2, d_s/2]$ along the three axis, where d_s is the control point spacing. This avoids self-folding and constrains the FFD to be a diffeomorphism over the volume. Additionally we repeat the deformation optimization at each control scale until it stabilizes to null displacements (in practice 4 iterations are generally sufficient) which proves more robust [12]. This also allows us to recover motions even larger than the maximal allowed in the coarsest scale.

5. Results

Since the problem of 3D occupancy grid probability updates from images has not yet been addressed, we first validate our solution, using various synthetic and real world datasets, shown in this section and the supplemental video. All of them are challenging for 4D analysis because of poorly textured surfaces, noise, outdoor lighting conditions, subject occlusion and object variety. Since ground truth motion and shape estimates are hardly accessible for real datasets, we provide a numerical analysis of the algorithm for synthetic datasets. Because existing 4D methods compute updates and flows of 2D manifolds, comparisons are difficult as they would require extracting and mapping a surface to each technique's particular surface estimates, which

introduces a bias. We will investigate this in future work. For all datasets, we use a 128^3 occupancy grid and three levels of control grid with control points 11, 7 and 3 voxels apart respectively. The EM converges in less than three iterations for all datasets. Our simple multi-threaded implementation takes less than 1 minute per frame for all datasets on an 8-core CPU. Vast speedups are still possible (GPU), as 99% of computation time is currently used to compute graph edge weights, and 1% for actual Fast-PD execution.

5.1. Synthetic Datasets

We generate motion sequences for a single ellipsoid in 9 known camera views. The algorithm is evaluated with different motions (translation and rotation), object shapes, and noise levels. The computed motion is examined against the ground truth motion in Fig. 4.

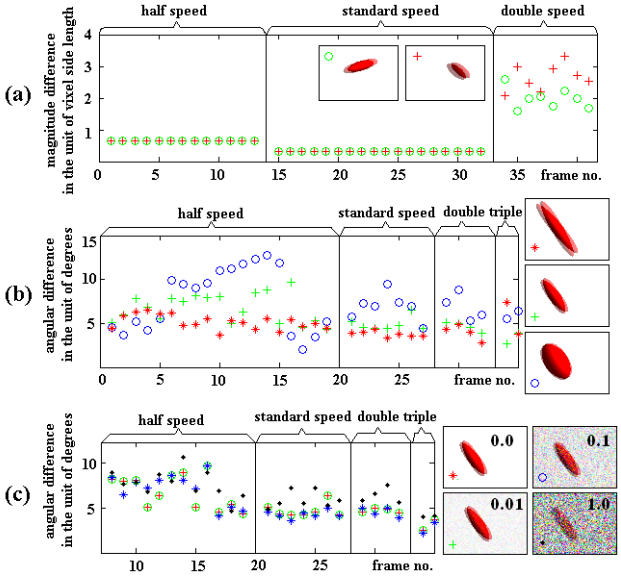

Figure 4. Synthetic dataset evaluation against the ground truth. Each plotted dot represents the average error of all voxels inside the ground truth ellipsoid at the frame. (a) translational motion along and orthogonal to the major principle axis direction with various speed; (b) rotational motion of three shapes with various angular speed; (c) rotational motion on image with 4 different noise level Gaussian noises. **Best viewed in color.**

For translational motion (Fig. 4(a)), two directions are tested with different velocity speeds. The motion field angular error is very close to zero. The average magnitude difference indicates that the algorithm performs best for ranges within the coarse control grid spacing. Small motions appear to be biased toward zero by smoothness constraints. Motions substantially bigger than the control grid maximum searching length are also not well retrieved.

For rotational motion (Fig. 4(b)), we plot the absolute 3D angular error (AAE). The ellipsoid on the top has twice as much polar radius as the middle. The bottom one has twice as much equatorial radii as the middle. The figure suggests that the angular motion should not to be too small. An interesting fact is that when the shape is close to a sphere, the motion ambiguity increases, and the algorithm is likely to fail, similar to the case of the optical flow of a textureless self-rotating sphere. Also in the future, a combination of silhouette and stereo inputs can be tested.

We evaluate response of the algorithm to Gaussian noise in images (Fig. 4(c)) in the case of rotation of the middle ellipsoid of Fig. 4(b) with different variances (0, 0.01, 0.1 and 1). Smaller noise levels yield smaller result errors as expected, but the method shows overall robustness, estimating reasonable motion solutions even in very noisy frames. A last synthetic dataset with multiple objects moving in the scene is also tested as shown in Fig. 3(a) and supplementary video, illustrating the capability of the algorithm to handle multiple object motions.

5.2. Real Datasets

Indoor BABY & SPIDER [4] and DANCER [24] are acquired with 8 camcorders running at $30fps$. Both BABY and SPIDER provide manually segmented silhouettes, but are otherwise inherently challenging. This illustrates the ability of the method to retrieve shape and motions for arbitrary shapes and rapid, self-occluded motions such as the spider legs (Fig. 3). Indoor ROND [11] includes walking and hand waving motion patterns. It is captured using 8 camcorders at $15fps$. Due to this relatively low frame rate, motions between frames are relatively large (some larger than 5 voxels), but the iterative multi-scale solution we propose recovers large motions correctly. Fig. 1 and Fig. 3(2) show the analysis of waving and walking motions in the sequence respectively. The motion tracks are computed by following the computed pairwise motion fields to track the history of some final voxels. The piece-wise linear effect of the motion track is not an artifact of our computation, but shows actual steps between frames, resulting from the combination of strong arm motion and relatively low video capture frame rate. Fig. 5 shows two slices of the occupancy grid at two time instants in the waving sub-sequence. The motion fields are overlaid on the probability slices, showing the dense nature of the result.

Outdoor SCULPTURE [13, 14] is acquired with 6 camcorders running at $30fps$. The camcorders are not color calibrated. There are sun light changes, shadows, reflections on the metallic sculpture. Given the noisy data and static background color models used, the computed occupancy grid shown includes high voxel probabilities for unwanted shapes, such as a shadow volume on the ground or sculpture, as visible in Fig. 3(3). Nevertheless, our framework is able to recover a coherent shape and dense flow estimate in spite of these underlying geometric incoherences. Such perturbations are likely to lead to failure of boundary-

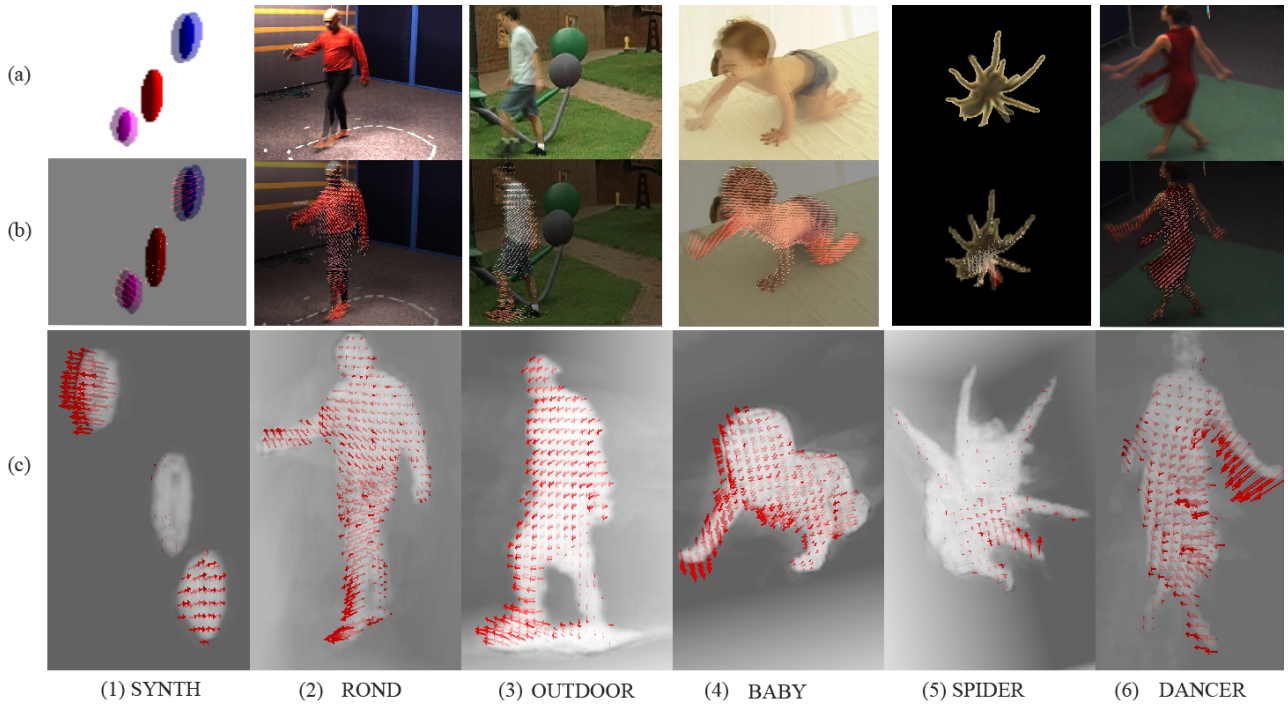

Figure 3. 3D occupancy flow result (**Best viewed in color**). Six datasets are illustrated column-wise. (a) The overlaid consecutive frames to indicate the motion; (b) the motion field on top of the respective image in (a); (c) the motion field and the occupancy probability grid from a novel view. Fig. 1 is part of the ROND dataset. Please check the supplementary video for the complete sequence results.

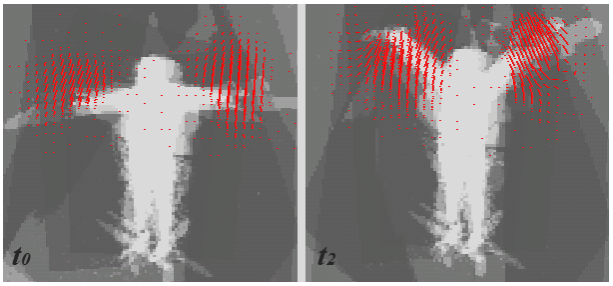

Figure 5. **Best viewed in color.** Occupancy slices at the same position in the volume at two time instants t_0 and t_2. The field is computed on the entire volume. The hands are waving up during the interval. 3D views are shown in Fig. 1.

based methods such as mesh tracking approaches. Fig. 6 shows the potential benefits of jointly estimating shape and motion to improve shape estimates. If we assume perfect silhouettes are available at time t_{258} (manually segmented for the purpose of the experiment), we can help the occupancy estimation at t_{259} and further. We apply our estmation to frames t_{258} and t_{259}. Occupancy probabilities at t_{258} act as a per-voxel prior, and clean the shadow region and reflection for t_{259} and later frames without using additional appearance models, e.g. for shadows. This suggests the potential for shape refinement across time of the proposed method, and the possibility of tracking and refining a probabilistic shape template while estimating dense 3D motions.

5.3. Applications

Besides shadow removal in noisy outdoor scenes, since the method can be applied to raw data without further assumptions, it can be used to retrieve motion segmentations and kinematic characteristics of the observed objects. We illustrate this with ROND and BABY sequences, by applying a simple EM clustering algorithm to simultaneously retrieve the number and parameters of sections in rigid motion and their corresponding voxels, using the displacement fields computed by our method, as shown in Fig. 7. This shows the potential for future that such characteristics can be extracted from the representation, which could in turn be reused to improve the shape estimation and motion tracking.

6. Discussion

We have explored a new direction in dense geometric and temporal 3D analysis, and propose a low-level approach to the problem of temporal updates of 3D occupancy probability grids. Initial applications have been explored in the case of silhouette inputs, but other input types (stereo, depth) can easily be included in the measurement term $\mathbf{\Phi}(\mathcal{G}_X)$. Experiments show the viability and robustness of the approach

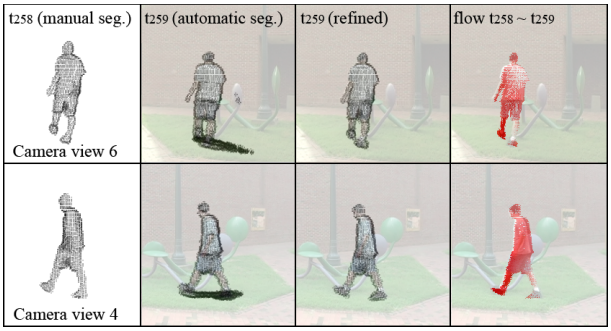

Figure 6. Occupancy refinement application. Column 1: the clean t_{258} grid computed from manually segmented silhouettes. Column 2: automatically estimated occupancy at t_{259}. Shadows and reflections are erroneously included due to naive automatically trained appearance models. Column 3: refined occupancy at t_{259} using the clean grid at t_{258} and the computed motion field between t_{258} and t_{259} of Column 4. All the occupancy and motion vectors are only plotted for points above probability 0.98. Column 2, 3 and 4 are overlaid with images at t_{259}. **Best viewed in color.**

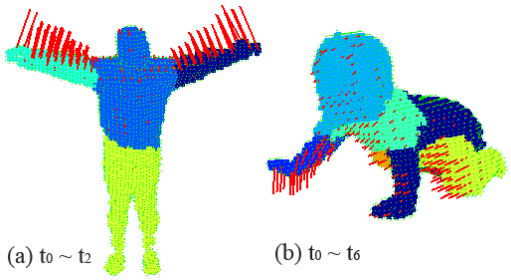

Figure 7. Rigid motion segmentation results (**best viewed in color**). Only voxels above probability 0.98 are used for this computation. (a) ROND result using motion from t_0 - t_2; (b) BABY result using motion from t_0 - t_6. Each rigid part is shown with a specific color. The rigid motion (translation and rotation) of each part is shown in red vectors, which explains the real motion at the time instant properly.

with various real datasets, and outdoor conditions challenging for stereo and surface-based methods. The method is promising and opens new possibilities and applications for motion segmentation with no geometric model or prior, or 3D tracking, kinematic structure inference, shape estimation, as our results show. Existing shape modeling and tracking methods could use our resulting fields as a cue to replace current 2D optical flow or sparse match inputs without having to explicitly deal with occlusion-related problems associated to an explicit boundary model. New temporal shape refinement schemes could be explored by using soft shape priors or using more past observations.

Acknowledgments: This work was partially supported by David & Lucille Packard Foundation Fellowship, the European Research Council grant 4DVIDEO and ANR-06-MDCA-003-01 DALIA project.

References

[1] J. S. D. Bonet and P. A. Viola. Roxels: Responsibility weighted 3d volume reconstruction. In *ICCV*, volume I(418-425), Sept. 1999.
[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *PAMI*, 2001.
[3] A. Broadhurst, T. Drummond, and R. Cipolla. A probabilistic framework for the Space Carving algorithm. In *ICCV*, 2001.
[4] G. Brostow, I. Essa, D. Steedly, and V. Kwatra. Novel skeletal representation for articulated creatures. In *ECCV*, 2004.
[5] R. L. Carceroni and K. N. Kutulakos. Multi-view scene capture by surfel sampling: From video streams to non-rigid 3d motion, shape and reflectance. *IJCV*, 49(2-3):175–214, 2002.
[6] G. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *CVPR*, June 2003.
[7] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. In *SIGGRAPH*, 2008.
[8] E. de Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel. Marker-less deformable mesh tracking for human shape and motion capture. In *CVPR*, 2007.
[9] F. Dellaert. The expectation maximization algorithm. Technical report, College of Computing, Georgia Institute of Technology, 2002.
[10] A. Elfes. Using occupancy grids for mobile robot perception and navigation. *IEEE Computer, Special Issue on Autonomous Intelligent Machines*, 22(6):46–57, June 1989.
[11] J.-S. Franco and E. Boyer. Fusion of multi-view silhouette cues using a space occupancy grid. In *ICCV*, 2005.
[12] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, and N. Paragios. Dense image registration through MRFs and efficient linear programming. In *Medical Image Analysis*, 2008.
[13] L. Guan, J.-S. Franco, and M. Pollefeys. 3d occlusion inference from silhouette cues. In *CVPR*, 2007.
[14] L. Guan, J.-S. Franco, and M. Pollefeys. Multi-object shape estimation and tracking from silhouette cues. In *CVPR*, 2008.
[15] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? In *PAMI*, 2004.
[16] N. Komodakis and G. Tziritas. Approximate labeling via graph-cuts based on linear programming. In *PAMI*, 2007.
[17] N. Komodakis, G. Tziritas, and N. Paragios. Fast, approximately optimal solutions for single and dynamic mrfs. In *CVPR*, June 2007.
[18] A. Laurentini. The Visual Hull Concept for Silhouette-Based Image Understanding. *PAMI*, 16(2):150–162, Feb. 1994.
[19] S. Lazebnik, Y. Furukawa, and J. Ponce. Projective visual hulls. *IJCV*, 74(2):137–165, 2007.
[20] J.-P. Pons, R. Keriven, and O. Faugeras. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *IJCV*, 72(2):179–193, 2007.
[21] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, pages 519–528, 2006.
[22] J. Starck and A. Hilton. Correspondence labeling for wide time frame free-form surface matching. In *ICCV*, 2007.
[23] D. Sun, S. Roth, J. Lewis, and M. Black. Learning optical flow. In *ECCV*, 2008.
[24] K. Varanasi, A. Zaharescu, E. Boyer, and R. Horaud. Temporal surface tracking using mesh evolution. In *ECCV*, 2008.
[25] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. *PAMI*, 27(3):475–480, 2005.
[26] S. Vedula, S. Baker, S. Seitz, and T. Kanade. Shape and motion carving in 6d. In *CVPR*, volume 2(2592), June 2000.
[27] D. Vlasic, I. Baran, W. Matusik, and J. Popovic. Articulated mesh animation from multi-view silhouettes. In *SIGGRAPH*, 2008.
[28] A. Wedel, C. Rabe, T. Vaudrey, T. Brox, and D. Cremers. Efficient dense 3D scene flow from sparse or dense stereo data. In *ECCV*, oct 2008.