Complete Multi-View Reconstruction of Dynamic Scenes from Probabilistic Fusion of Narrow and Wide Baseline Stereo

Tony Tung Shohei Nobuhara Takashi Matsuyama
Graduate School of Informatics, Kyoto University, Japan
{tung,nob,tm}@vision.kuee.kyoto-u.ac.jp

Abstract

This paper presents a novel approach to achieve accurate and complete multi-view reconstruction of dynamic scenes (or 3D videos). 3D videos consist in sequences of 3D models in motion captured by a surrounding set of video cameras. To date 3D videos are reconstructed using multi-view wide baseline stereo (MVS) reconstruction techniques. However it is still tedious to solve stereo correspondence problems: reconstruction accuracy falls when stereo photo-consistency is weak, and completeness is limited by self-occlusions. Most MVS techniques were indeed designed to deal with static objects in a controlled environment and therefore cannot solve these issues. Hence we propose to take advantage of the image content stability provided by each single-view video to recover any surface regions visible by at least one camera. In particular we present an original probabilistic framework to derive and predict the true surface of models. We propose to fuse multi-view structure-from-motion with robust 3D features obtained by MVS in order to significantly improve reconstruction completeness and accuracy. A min-cut problem where all exact features serve as priors is solved in a final step to reconstruct the 3D models. In addition, experimental results were conducted on synthetic and challenging real world datasets to illustrate the robustness and accuracy of our method.

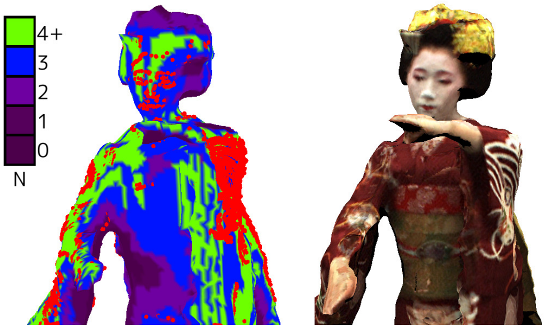


Figure 1. **Complete multi-view reconstruction**. Stereo correspondence is challenging if the number of input images N is low. Left: Red dots represent 3D points with high stereo photo-consistency. The region around the chest is not visible by many cameras ($N \leq 3$) and has low stereo photo-consistency. Right: The true surface is recovered using temporal cues from narrow baseline stereo.

1. Introduction

Multiple view stereo reconstruction of dynamic scenes (or 3D video) is an imaging technique which consists in sequences of 3D models captured by a surrounding set of calibrated and synchronized video cameras [18, 21, 7, 5, 2, 27]. It produces free-viewpoint videos of one or several models in motion in an immersive environment. The technology requires no special equipment to wear such as a suit with markers. As subjects are free to perform any actions, this system fits to a very wide range of motion capture applications: cultural heritage preservation (e.g. traditional dance recording), medicine, sports, entertainment, and so on.

To date, due to hardware arrangement, existing systems have used multi-view wide baseline stereo (MVS) reconstruction algorithms to produce 3D videos. Every frame is reconstructed independently. Nevertheless to obtain very accurate 3D models from stereo a high number of input views are required. While the reconstruction of static objects returns very good results [24], in the 3D video framework, only ten to twenty video cameras are usually available, frame resolution is low, and lighting is not controlled. 3D models of humans in action usually present poorly photo-consistent and self-occluded regions which are challenging to reconstruct. Consequently, the 3D video reconstruction process is still open to many problems concerning accuracy and completeness. As a matter of fact, although MVS provides accurate depths (with large triangulation angles), it suffers from correspondence and occlusion issues since viewpoints are mainly different or insufficient in some respect. On the other hand, correspondences become easy to estimate and occlusions are less frequent with narrow baseline stereo since viewpoints are very similar [6, 15].

We propose to introduce temporal cues into 3D video to overcome the limitations of MVS reconstruction. Narrow baseline stereo can provide dense reconstruction up to a scale using structure-from-motion techniques (cf. Figure 1). Thus we fuse 3D structures computed from multi-view optical flows to robust 3D features computed from MVS. Both techniques are complementary since narrow baseline stereo can find information where the wide baseline stereo cannot (e.g. in regions visible by only one camera), and wide baseline stereo helps to recover the scale factor of 3D structures. The fusion is formulated in a novel Bayesian probabilistic framework to estimate the true surface of models. Finally, a min-cut problem is formulated and efficiently solved using graph-cuts. Furthermore experimental results were conducted on synthetic and real data and focus on the reconstruction of poorly textured regions where MVS reconstruction fails due to insufficient number of input images.

The rest of the paper is organized as follows. The next section discusses work related to the techniques presented in this paper. Section 3 presents the 3D feature extraction from narrow and wide baseline stereo. Section 4 gives a description of the new probabilistic fusion scheme. Section 5 presents the min-cut formulation problem. Section 6 shows experimental results. Section 7 concludes with a discussion on our contributions.

2. Related work

Since a decade an increasing number of immersive stereo capture systems have become available [18, 21, 7, 5, 2, 27] (cf. Figure 2). To date, multi-view wide baseline stereo (MVS) reconstruction techniques have been used to reconstruct the 3D video frames (see [24] for a survey). Using MVS implies indeed to cope with challenging stereo correspondence problems. Due to the complexity of the 3D video framework, reconstruction completeness is usually poor. On the other hand, the models are reconstructed independently. Reconstruction errors are not propagated in the whole sequence and topology of models can vary freely.

Several methods were proposed to obtain dense reconstruction of non-rigid models with various constraints. For example in [16] the reconstruction is performed from a single camera with multispectral lighting, and in [10] models are reconstructed from multiple video streams by tracking 3D features over time. These approaches return impressive results but special clothes are necessary, and reconstructions consist in a deforming surface with fixed topology for the whole sequence.

Alternatively, many methods combining spatial and temporal cues have been proposed in the literature. In [13], the problem is formulated as a 3D weighted hypersurface embedded in space-time. The optimal model is obtained by minimizing an energy functional following photo-consistency criteria. Nevertheless the temporal constraint is purely based on the geometric smoothness of the 4D surface and does not employ temporal photo-consistency. In [33], shape, motion and in particular 3D motion flows (i.e. scene flows [32]) are computed simultaneously. The scene is represented at two consecutive time instants in a 6D space to carve. The method relies on a time-space photo-consistency computation to extract accurate 3D features. As well, [35] computes shape and motion simultaneously based on photometric constraints yet assuming rigid motions and no occlusions. In the 3D video framework, it sounds natural to combine spatial and temporal cues in order to reconstruct the sequences. However a strong dependency to temporal cues is somehow a consistency constraint which limits the surface evolution over time. Consequently, applying MVS on every frame independently is still the most robust approach to handle topology changes and self-occlusions. In [31], a super-resolution approach is proposed to increase the accuracy of 3D video reconstruction. Although improvements can be observed locally, this approach requires several cameras to compute super-resolution, and therefore cannot improve significantly the reconstruction completeness. Let us mention that a wide range of papers aims to recover depth maps from stereo arrangement using fusion of stereo with motion (or X) [8, 36]. However the extension to complete multi-view reconstruction is not straightforward since occlusions have to be managed.

In this paper, we propose a probabilistic framework to derive an adaptative ballooning term as [17] in order to formulate a min-cut problem to recover 3D surfaces [3, 19, 30, 27, 34]. Using probabilistic approach allows indeed to soften constraints and recover partially occluded views. In previous work, all existing probabilistic frameworks were designed to recover visibility using stereo

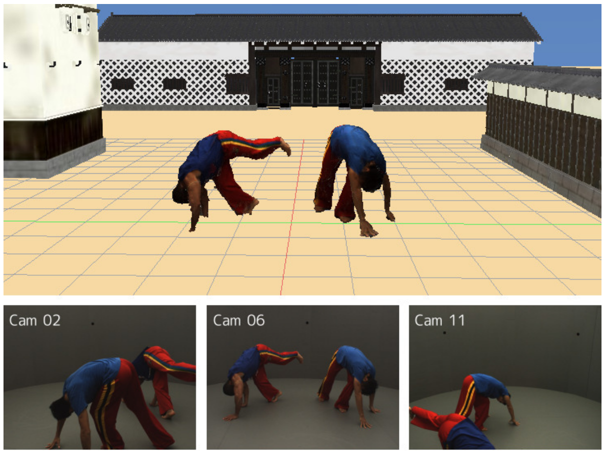


Figure 2. **3D video.** Top: Multi-view reconstruction from 14 video cameras. Bottom: Corresponding video frames. Stereo correspondence is challenging as several regions can be occluded in many views.

photo-consistency criteria [1, 4, 28, 11, 12, 17, 22]. Hence every region has to be seen by at least two views. However this cannot be guaranteed in the 3D video framework and leads to poor results. To our knowledge, we are the first to propose a probabilistic fusion of multi-view structure-from-motion with robust MVS features to reconstruct dynamic scenes (as most of previous work propose depth map fusion). Our approach allows to recover surface regions visible by only one camera, and thus the completeness is significantly improved.

3. Complete 3D feature recovery

Multi-view wide baseline stereo allows to accurately reconstruct static 3D objects. In particular, very good results can be obtained with a robust stereo photo-consistency criterion [24]. Nevertheless the performance are limited in the 3D video framework due to the low number of cameras (some regions are seen only by one camera), video frame resolution, lighting variations and motion noise: dense correspondence finding is very challenging. On the other hand, image contents and lighting are consistent for consecutive frames captured at video frame rate from single viewpoints. Assuming piecewise rigidity, even the regions seen by only one camera can be recovered using motion fields. Hence reconstruction completeness can be significantly increased by fusion of both narrow and wide baseline stereo features.

3.1. Features from narrow baseline stereo

Consecutive frames from single-view video contain lots of similarities. Many correspondences can be found efficiently using a feature-based tracker (e.g. KLT [26], SIFT [20], DAISY [29]). Having sufficient pairs of corresponding points in two images allows to recover the feature positions in the 3D space (up to a scale factor) using structure-from-motion techniques [14, 6, 15]. Nonetheless this result is valid only under rigidity assumption. In our framework we cope with human models in motion. We believe a rigidity assumption is locally reasonable even with loose clothing due to the acquisition frame rate (25fps). Hence we propose to group optical flows into clusters based on the following simple criteria: length, orientation and position. This can be robustly achieved using normalized spectral clustering [25]. The correlation function $\rho_{i,j}$ which serves to construct the similarity matrix (and consequently a graph Laplacian matrix) between the trajectory of two feature points i and j is defined as:

$$\rho_{i,j}^2 = \omega_l D_l^2(i,j) + \omega_\theta D_\theta^2(i,j) + \omega_p D_p^2(i,j), \quad (1)$$

where D_l, D_θ, D_p are normalized distances for optical flow length, angle and position respectively, and $\omega_l = 0.7$, $\omega_\theta = 0.2$, $\omega_p = 0.1$ are empirically determined weighting factors. The number of clusters K is set to 30. If K is overestimated then the flow field will be overpartionned, but without any impact on the quality of the reconstruction. 3D structures are then recovered from every optical flow cluster (cf. Figure 3)

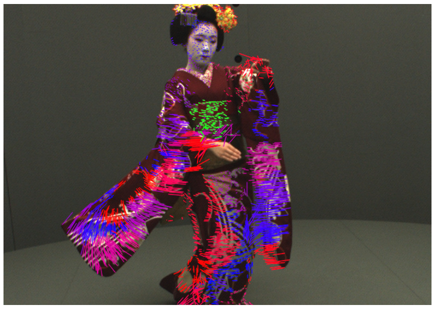


Figure 3. **Motion flow clusters.** Here, 2941 motion features were detected and grouped in 30 clusters. Spectral clustering is used to cluster optical flows according to length, orientation and position. Piecewise rigidity can be assumed between consecutive frames due to the video frame rate (25fps).

3.2. Features from multi-view wide baseline stereo

Robust stereo photo-consistent 3D features are extracted from multiple view wide baseline images using an approach inspired from [9]. However a more selective constraint has been added to ensure uniqueness of features: a feature is valid if and only if it has a high stereo photo-consistency score and no other candidate exists on its epipolar line. Thus we obtain a set of robust 3D features for every 3D video frame (cf. Figure 4 Left).

4. Probabilistic fusion of features

We propose an original probabilistic fusion framework to estimate the true surface of models. Our formulation involves features computed by both narrow and wide baseline stereo techniques (cf. Section 3). It can be expressed as a maximum a posteriori Markov Random Field (MAP-MRF) problem and solved using global optimization algorithms [28, 30]. The posterior probability to maximize is:

$$p(\Theta|\Phi,\Gamma) \propto \prod_i E_p(\theta_i,\phi_i) E_q(\theta_i,\gamma_i) \prod_i \prod_{j\in\mathcal{N}(i)} V(\theta_i,\theta_j), \quad (2)$$

where $\Theta = \{\theta_i\}$ denotes the 3D surface points at t (hidden parameters), $\Phi = \{\phi_i\}$ represents the input images at t (observed data), $\Gamma = \{\gamma_i\}$ represents the structures found from motion flows at t (observed data), E_p is the local evidence for node i based on the robust stereo photo-consistency score estimated at Θ from Φ, E_q is the local evidence for node i being on the true surface, $N(i)$ represents the neighbors of node i, and V is a smoothness assumption.

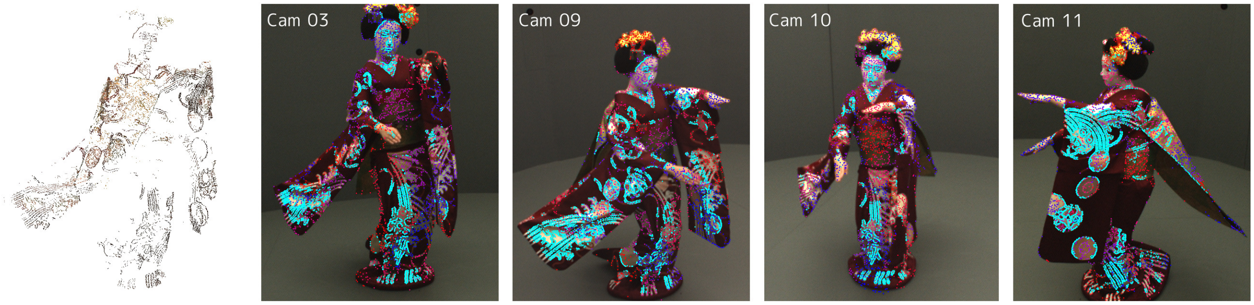


Figure 4. **Feature fusion.** Robust 3D features are extracted by multi-view wide baseline stereo (cf. Left), and projected into image planes where photo-consistency is strong (in cyan). If the 3D features fall into a region where motion features exist, then 3D structures can be estimated with scale factors. Note that robust stereo fails to extract features on the chest whereas many single-view motion features are detected.

Our probabilistic fusion scheme aims indeed to recover the unknown scale factors $\Sigma = \{\sigma_i\}$ of 3D structures Γ from the estimated sparse features Φ^* in order to obtain the surface points Θ^* belonging to the true surface:

$$\Theta^* = \arg\max_{\Theta} p(\Theta|\Phi, \Gamma) = \arg\max_{\Theta} \int_{\Sigma} p(\Phi, \Sigma, \Phi^*, \Gamma) d\Sigma. \quad (3)$$

Σ is used as a set of discrete labels to represent the state of each hidden variables. Thus the true surface points Θ^* can be expressed as Eq. 3, where the most probable values of Θ have to be found, given the observed data Φ and Γ, and marginalizing out Σ. The problem is solved using graph-cuts as a trade-off of accuracy and speed (cf. Section 5).

4.1. Problem initialization

Let assume $Q_i = \{(\mathbf{x}_i^j, \sigma_i)\}$ is the set of points defined by the 3D coordinates $\mathbf{x}_i^j$ and the scale factor σ_i, and let denote $\mathcal{S}$ the (yet unknown) true surface to reconstruct at t, and $\mathcal{V}$ the visual hull computed from the M image silhouettes at t. Our algorithm starts by approximating $\mathcal{S}$ with $\mathcal{V}$ using a best fit in the least-squares sense. Let σ_i^0 be the initial value of σ_i at t_0:

$$\sigma_i^0 = \arg\min_{\sigma_i} \sum_{q \in Q_i} (q - f_{\mathcal{V}}(q))^2, \quad (4)$$

where $f_{\mathcal{V}}$ is a function that projects a 3D point onto the surface $\mathcal{V}$ with respect to its corresponding image viewpoint. The probability of a point $\mathbf{x}$ lying on $\mathcal{V}$ should reflect the uncertainty on whether or not $\mathbf{x}$ belongs to $\mathcal{S}$.

Let denote Φ^* the set of sparse robust 3D features computed at t by multi-view wide baseline stereo as described in Section 3.2. In theory, $\forall \mathbf{x} \in \Phi^* : \mathbf{x} \in \mathcal{S}$. Assuming a reference image ϕ_r, and $P_r \subset \Phi^*$ the set of features visible[1] in ϕ_r, if there exists a subset $P_r^j \subset P_r$ which projection into ϕ_r lies in a 2D feature cluster envelope C_j (as defined in Section 3.1), then the 3D structures Q_j derived from C_j are scaled to fit P_r^j in the least-squares sense:

$$\sigma_j^0 = \arg\min_{\sigma_j} \sum_{q \in Q_j} (q - P_r^j)^2, \quad (5)$$

where Q_j is the set of 3D points having the scale factor σ_j. The probability of a point $\mathbf{x}$ lying near a robust 3D feature should reflect how well $\mathbf{x}$ is close to the true surface $\mathcal{S}$.

4.2. Feature probabilistic evidence

In this section we formulate the evidence E_q as an explicit estimation of the probability that a 3D scene point lies on the true surface $\mathcal{S}$. It relies on the fact that the sparse 3D features are assumed to lie on $\mathcal{S}$, and therefore if their projections $\{p_j\}$ into the input images Φ coincide with extracted motion features $\{c_j\}$, then the 3D structures $\{\gamma_j\}$ found from $\{c_j\}$ can be recovered with the right scale factors. In this case $\{\gamma_j\} \in \mathcal{S}$. Thus we define the probability evidence E_q as:

$$E_q = f(\Delta), \quad (6)$$

$$\Delta = \min_{p \in \{p_j\}, q \in \{c_j\}} (\|p - q\|), \quad (7)$$

where the function $f \in [0, 1]$ satisfies the following properties:

1. if Δ is small, $\{\gamma_j\}$ is close to $\mathcal{S}$, then f is high.

2. f decreases as Δ increases and conversely.

The probabilistic evidence for a 3D point to lie on the true surface is estimated for each structure set in every image. In our experiments, f is defined as a Gaussian distribution centered on the 3D surface and truncated in the inside part of the object.

[1] The visibility can be derived from the visual hull.

4.3. Feature fusion

Assuming a reference image ϕ_r and the 3D structures $\{\gamma_r\}$ found from motion features $\{c_r\} \in \phi_r$, let denote E_r the evidence for $\{\gamma_r\}$ being on $\mathcal{S}$, and $R \subset \mathcal{V}$ the surface region where the set $\{c_r\}$ projects onto. Using the visual hull, we derive the image views $\{\phi_i\}_{i \neq r}$ where R is visible, extract the motion features $\{c_i\}_{i \neq r} \in \{\phi_i\}_{i \neq r}$ falling into R, and then apply the process described in Section 4.1 and 4.2 to estimate the evidence $\{E_i\}_{i \neq r}$ for the 3D structures $\{\gamma_i\}_{i \neq r}$ (found from $\{c_i\}_{i \neq r}$) to be on $\mathcal{S}$.

Finally, the N largest local evidences $\{E_i\}_{i \in [1,N]}$ are aggregated to obtain the evidence E_q for the points $\bigcup_{i=1}^{N} \{\gamma_i\}$ to be on $\mathcal{S}$:

$$E_q = \prod_{i=1}^{N} E_i, \tag{8}$$

where $\forall i, E_i > E_{i+1}$. Point sets having too low evidences are discarded from the reconstruction process.

4.4. Temporal cues

Let denote $p(Q_t)$ the probability for $Q_t = \{(\mathbf{x}_t^j, \sigma_t)\}$ to be surface points of the model reconstructed at t. According to Bayes' theorem, the posterior probability $p(Q_{t+1})$ at $t+1$ can be estimated using $p(Q_t)$ as prior probability:

$$p(Q_{t+1}) = \frac{p(Q_t) E_q(Q_t)}{p(Q_t) E_q(Q_t) + (1 - p(Q_t))}. \tag{9}$$

Consequently, the probability for 3D points Q_{t+1} to have scale factor σ_t is estimated to $p(Q_{t+1})$ at $t+1$.

5. 3D shape reconstruction

Labeling cost from 3D feature observations. The probabilistic evidence presented in Section 4 is used as an intelligent ballooning term $i(\mathbf{x})$ as introduced in [17]. However as explained in Section 2, unlike all previous work involving a probabilistic framework, our formulation does not aim to recover visibility using stereo photo-consistency criteria, but rather estimate and predict true model surfaces from 3D feature observations.

Min-cut problem formulation. The 3D model reconstruction is defined as a minimization of a Markov Random Field (MRF) energy. A volumetric graph-cut is formulated where 3D features serve as priors [3]. We map 3D voxels in the scene as graph nodes, and set the weights of s- and t-links of each node to $i(x)$ while n-links store photo-consistencies. 3D features Φ are embedded as *definitely object surface points* as proposed in [30, 27]. Local constraints are crucial to prevent shrinking of thin parts. This approach is well known to converge to an accurate solution in a polynomial time.

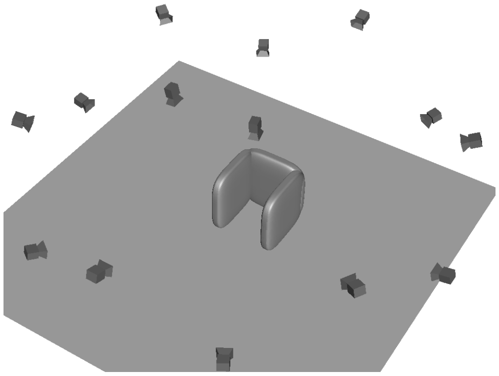


Figure 5. **Synthetic 3D video with camera arrangement.**

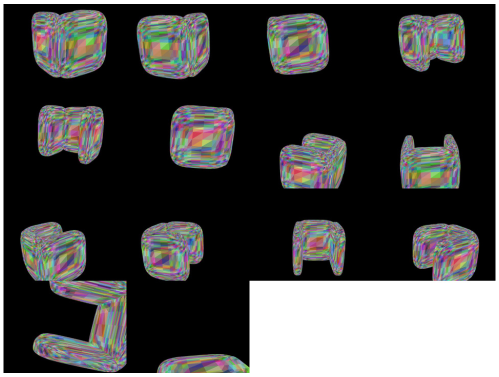


Figure 6. **Multiple views of the synthetic model with texture.**

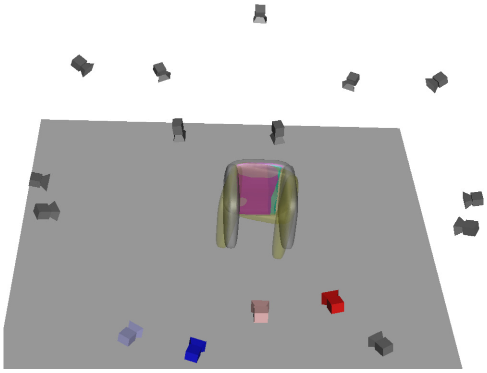
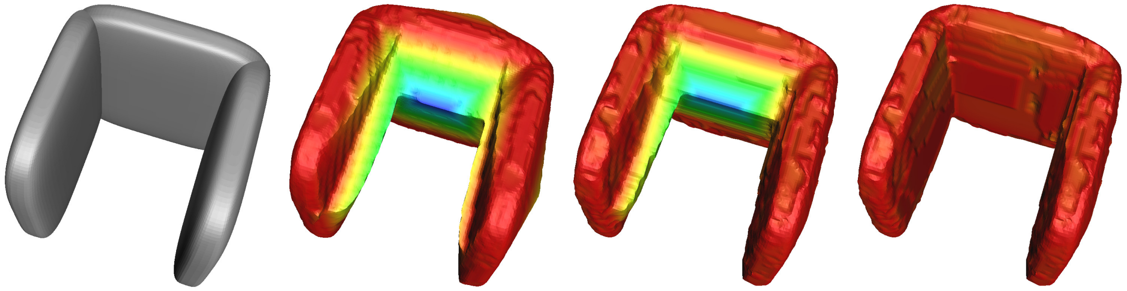


Figure 7. **3D structure from narrow baseline stereo.** 3D surfaces (in cyan and magenta) in the concave region are estimated at t from real and virtual camera views (in light blue and light red, and in dark blue and dark red respectively).

6. Experimental results

To evaluate the performance of our approach, we have tested our algorithm on synthetic and real 3D video sequences. The same camera arrangement was used for all datasets (cf. Figure 5). The framework consists in 14 XGA video cameras located at ∼3m distance from the center of



(a) Ground truth. (b) Visual hull. (c) MVS [30, 27]. (d) Proposed.

Figure 8. **3D model reconstruction comparison.** The colors indicate the distance to the ground truth surface (red is closer). The concavity is not well recovered by (b) and (c), whereas (d) performs well. Distances are computed using [23].

the studio. Frame rate is 25fps. Computation time takes 10min to generate one model using an Intel Core2Duo 3.0GHz with 1cm voxel grid resolution.

Synthetic sequence. The size of the synthetic model is $1\text{m} \times 1\text{m} \times 1\text{m}$. The ground truth is represented in Figure 5. All multi-view frames of the model with texture are shown in Figure 6. The synthetic model includes a highly concave region which depth and geometry are difficult to recover using classical multi-view wide baseline stereo techniques (MVS). However using our probabilistic fusion scheme we show that 3D structures from the concavity can be accurately recovered. Figure 7 shows the 3D surface recovered by temporal narrow baseline stereo from two successive captures by two camera views. The model is represented in gray at time t and is seen by the light blue and light red cameras. The model in motion is represented in yellow at $t + 1$. The equivalent motion (translation and rotation) can be applied to the cameras at t: the dark blue and dark red cameras denote the virtual positions of the real light blue and light red cameras at $t + 1$ respectively. At time t, the light blue and dark blue cameras estimate the 3D surface in cyan, and the light red and dark red estimate the 3D surface in magenta using the approach described in this paper. Figure 8 shows: (a) the ground truth, (b) the visual hull, (c) the 3D model estimated from wide baseline MVS [30, 27] and (d) the 3D model estimated by the proposed method. As the concave region is mostly seen by only one single viewpoint, it cannot be completely reconstructed using wide baseline MVS only (without temporal cues). In Figure 8, (c) and (d) clearly show the improvement by our proposed method. As well, quantitative evaluations were reported in Table 1. *Accuracy* is the distance d (in cm) such that 90% of the reconstructed surface is within d cm of the ground truth, and *completeness* measures the percentage of the reconstructed surface that are within 2 cm of the ground truth [24]. For $d = 2$cm, the *accuracy* measure returns 56.77%, 83.26% and 98.20% for visual hull, MVS [30, 27] and our proposed method respectively. The probabilistic fusion of narrow and wide baseline stereo features outperforms MVS on the synthetic dataset.

	Accuracy	Completeness
Visual hull	7.73cm	80.20%
MVS [30, 27]	5.09cm	86.73%
Proposed	1.32cm	90.15%

Table 1. **Reconstruction accuracy and completeness**.

Real sequences. Two challenging real sequences illustrate this paper. Both *maiko* sequence and *capoeira* sequence contain difficult regions to recover due to insufficient visibility from camera views and lack of texture (cf. Figure 1). For example, the *maiko* dances slowly and has long sleeves which often occlude her chest in many camera views. As well, the *capoeira* performer moves his arms relatively fast in front of his chest, and wears a plain blue T-shirt which has no texture except from a logo. Nevertheless while wide baseline MVS fails to find features, narrow baseline can recover 3D structures from motion features thanks to image content stability between consecutive frames from single-view videos (cf. Figures 3, 4 and 11). Outlying optical flows were filtered using RANSAC with a hypothesized inlier ratio of 70%. Figures 9 and 12 illustrate failed reconstructions using [30, 27] against our results. In Figures 10 and 13 models are shown with close-ups on the chest region. Visual hulls contain phantom volumes due to silhouette projection ambiguities (top row). Regions having weak stereo photo-consistency are not well carved by MVS and protrusions remain (middle row). Finally our method is able to recover the self-occluded regions by estimating the true surface position, and produce smooth surfaces where rendered texture quality is clearly better than classic MVS re-

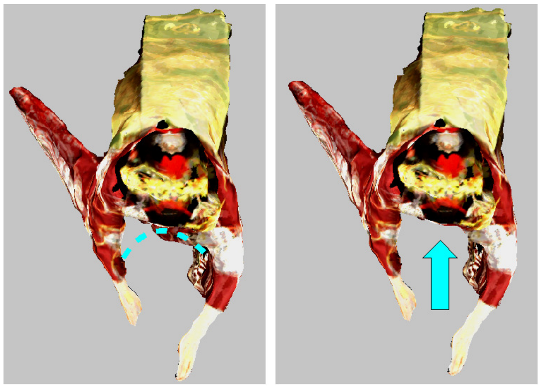


Figure 9. Left: MVS [30, 27]. Right: Proposed.

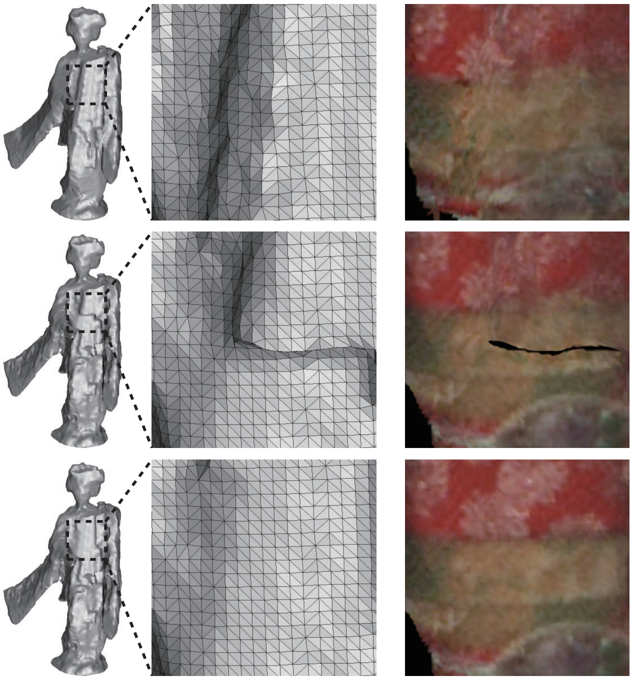


Figure 10. **Reconstruction results**. Top: Visual hull. Middle: MVS [30, 27]. Bottom: Proposed.

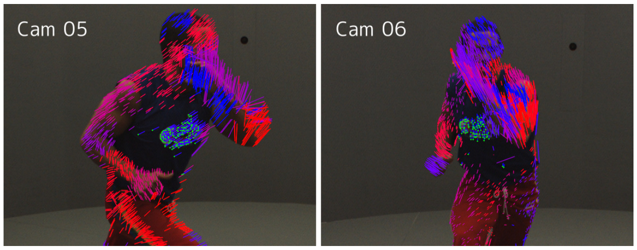


Figure 11. **Optical flow extraction between consecutive frames can be achieved taking advantage of image content stability.**

construction (bottom row). These results show that probabilistic fusion of narrow and wide baseline stereo features achieves accurate and complete dynamic 3D shape reconstruction.

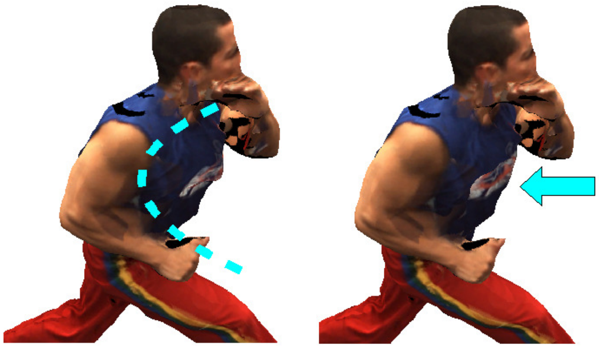


Figure 12. Left: MVS [30, 27]. Right: Proposed.

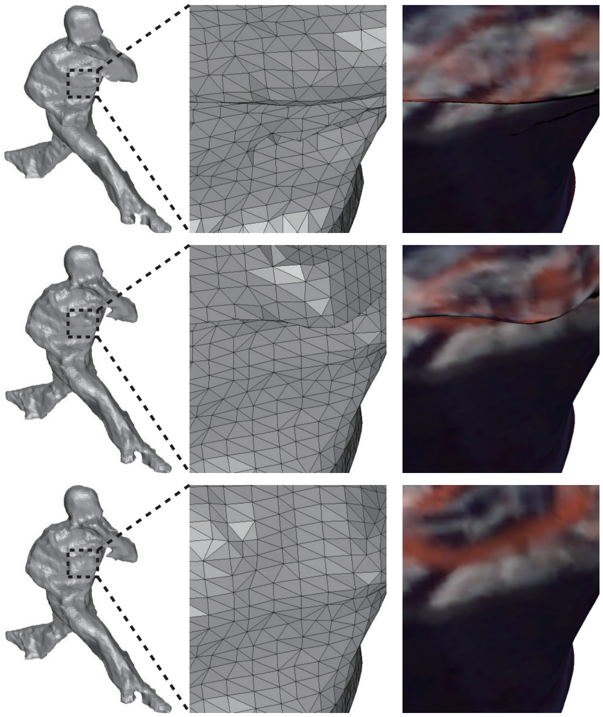


Figure 13. **Reconstruction results**. Top: Visual hull. Middle: MVS [30, 27]. Bottom: Proposed.

7. Conclusion

Multi-view reconstruction of dynamic scenes (3D video) faces many issues that cannot be solved with existing multi-view stereo (MVS) reconstruction approaches. In the 3D video framework, stereo correspondences are very challenging since sequences of 3D models in motion are captured by a lower number of cameras, frame resolution is lower, lighting is subject to variations, occlusions occur frequently, and motion noise induces inaccurate matching.

In this paper, we present a novel method to achieve accurate and complete multi-view reconstruction of dynamic scenes. Our approach consists in the probabilistic fusion of narrow and wide baseline stereo features to recover data where classical MVS reconstruction techniques fail. In particular, our method allows to recover partially occluded regions seen by only one camera (whereas previous work requires at least two or more input images). Thus the reconstruction completeness is consequently improved. The true surface of models are estimated using a probabilistic scheme involving 3D features having robust photo-consistency and 3D structure-from-motion derived from single views. Consecutive frames from single-view cameras show indeed sufficient consistency to take advantage of temporal cues such as motion features. Hence we can estimate and predict locally true surfaces using Bayesian inference. Finally, a global optimization of min-cut problem is solved using graph-cuts in order to reconstruct the complete 3D model sequences.

To our knowledge, our formulation is novel, and to date no existing method can solve the 3D video correspondence issues we pointed out. The performance of our approach is demonstrated on both synthetic and complex real world data.

Acknowledgments

This work was supported in part by the Japan Society for the Promotion of Science (JSPS No.20.08366), MEXT of Japan under the GCOE program "Informatics Education and Research Center for Knowledge-Circulating Society", and JST-CREST project "Foundation of Technology Supporting the Creation of Digital Media Contents".

References

[1] M. Agrawal and L. Davis. A probabilistic framework for surface reconstruction from multiple images. *CVPR*, 2:470–477, 2001.

[2] J. Allard, C. Ménier, B. Raffin, E. Boyer, and F. Faure. Grimage: Markerless 3d interactions. *SIGGRAPH - Emerging Technologies*, 2007.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2001.

[4] A. Broadhurst, T. Drummond, and R. Cipolla. A probabilistic framework for space carving. *ICCV*, 1:338–393, 2001.

[5] K. M. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette across time: Part ii: Applications to human modeling and markerless motion tracking. *IJCV*, 63(3):225–245, 2005.

[6] O. Faugeras, Q.-T. Luong, and T. Papadopoulo. The geometry of multiple images. *MIT Press*, 2001.

[7] J. Franco, C. Menier, E. Boyer, and B. Raffin. A distributed approach for real-time 3d modeling. *CVPR Workshop on Real-Time 3D Sensors and their Applications*, page 31, 2004.

[8] U. Franke, C. Rabe, H. Badino, and S. Gehrig. 6d-vision: Fusion of stereo and motion for robust environment perception. *DAGM*, 2005.

[9] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. *CVPR*, 2007.

[10] Y. Furukawa and J. Ponce. Dense 3d motion capture from synchronized video streams. *CVPR*, 2008.

[11] P. Gargallo and P. Sturm. Bayesian 3d modeling from images using multiple depth maps. *CVPR*, 2:885–891, 2005.

[12] M. Goesele, B. Curless, and S. Seitz. Multi-view stereo revisited. *CVPR*, 2:2402–2409, 2006.

[13] B. Goldlücke, I. Ihrke, C. Linz, and M. Magnor. Weighted minimal hypersurface reconstruction. *PAMI*, 29(7):1194–1208, 2007.

[14] R. Hartley. In defense of the eight-point algorithm. *PAMI*, 19(6):580–593, 1997.

[15] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. *Cambridge University Press*, 2003.

[16] C. Hernandez, G. Vogiatzis, G. J. Brostow, B. Stenger, and R. Cipolla. Non-rigid photometric stereo with colored lights. *ICCV*, 2007.

[17] C. Hernandez, G. Vogiatzis, and R. Cipolla. Probabilistic visibility for multi-view stereo. *CVPR*, 2007.

[18] T. Kanade, A. Yoshida, K. Oda, H. Kano, and M. Tanaka. A stereo machine for video-rate dense depth mapping and its new applications. *CVPR*, 1996.

[19] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. *ECCV*, pages 82–96, 2002.

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[21] T. Matsuyama, X. Wu, T. Takai, and S. Nobuhara. Real-time 3d shape reconstruction, dynamic 3d mesh deformation, and high fidelity visualization for 3d video. *CVIU*, 96(3):393–434, 2004.

[22] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J. Frahm, R. Yang, D. Nister, and M. Pollefeys. Real-time visibility-based fusion of depth maps. *ICCV*, 2007.

[23] R. S. P. Cignoni, C. Rocchini. Metro: measuring error on simplified surfaces. *Computer Graphics Forum, Blackwell Publishers*, 17(2):167–174, 1998.

[24] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. *CVPR*, 2006.

[25] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 22(8):888–905, 2000.

[26] J. Shi and C. Tomasi. Good features to track. *CVPR*, pages 593–600, 1994.

[27] J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 2007.

[28] C. Strecha, R. Fransens, and L. V. Gool. Wide-baseline stereo from multiple views: a probabilistic account. *CVPR*, 2004.

[29] E. Tola, V. Lepetit, and P. Fua. A fast local descriptor for dense matching. *CVPR*, 2008.

[30] S. Tran and L. Davis. 3d surface reconstruction using graph cuts with surface constraints. *ECCV*, pages 219–231, 2006.

[31] T. Tung, S. Nobuhara, and T.Matsuyama. Simultaneaous super-resolution and 3d video using graph-cuts. *CVPR*, 2008.

[32] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. *ICCV*, 1999.

[33] S. Vedula, S. Baker, S. Seitz, and T. Kanade. Shape and motion carving in 6d. *CVPR*, 2000.

[34] G. Vogiatzis, C. Hernandez, P. Torr, and R. Cipolla. Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *PAMI*, 29(12):2241–2246, 2007.

[35] L. Zhang, B. Curless, A. Hertzmann, and S. M. Seitz. Shape and motion under varying illumination: Unifying structure from motion, photometric stereo, and multi-view stereo. *ICCV*, 2003.

[36] J. Zhu, L. Wang, R. Yang, and J. Davis. Fusion of time-of-flight depth and stereo for high accuracy depth maps. *CVPR*, 2008.