Minimal 3D Video

Tony Tung [*] Takashi Matsuyama [†]
Graduate School of Informatics, Kyoto University, Japan

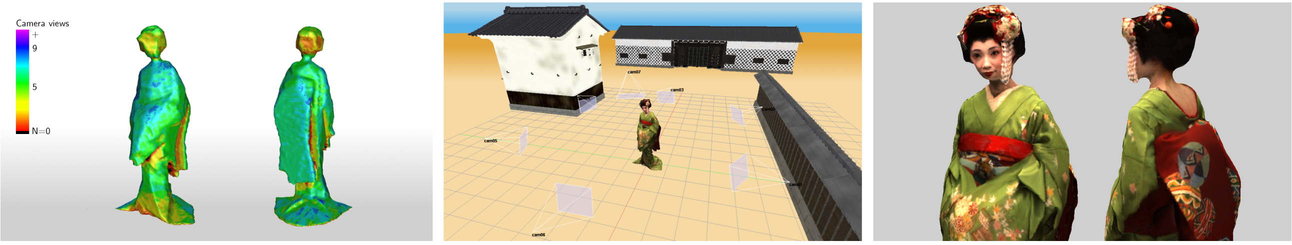

Figure 1: *(Left) The 3D model surface is colored with respect to the visibility from 15 camera viewpoints. Almost all the regions are visible by at least two camera ($N > 1$). (Center) 3D video frame and minimal 3D video setting with only 7 video cameras. (Right) 3D frame close-up.*

1 Introduction

We present a new concept that achieves the 3D reconstruction of dynamic scenes from multi-view video cameras (or 3D videos) using a minimal number of cameras, as opposed to the present state of the art approaches which require either several tens of cameras or high definition devices. A 3D video consists of a sequence of 3D models in motion captured by a surrounding set of video cameras. The result is a video where observers can choose freely their viewpoints. It is a markerless motion capture system where subjects do not need to wear special equipment. Hence, this system suits to a very wide range of applications (e.g. entertainment, medicine, sports, and so on). The 3D models are obtained using image-based multi-view stereo reconstruction techniques (or MVS). The performance of MVS relies on the quality and quantity of images taken from different viewpoints. As stereo correspondences have to be found between the images, the reconstruction fails in the case of weak stereo photo-consistency due to lack of camera views or lighting variations: consistent information is necessary.

To date, most 3D video systems have used shape reconstruction methods inspired from the 3D reconstruction of static scenes. In this case the 3D video frames are processed separately. The use of temporal cues derived from feature tracking is still limited to high definition setups (e.g. [de Aguiar et al. 2008] uses a 3D scan). As well, video cameras have to be geometrically calibrated, synchronized and linked to a PC cluster which processes the captured frames. The overall system is therefore expensive and tedious to manage. Nevertheless in the recent work of [Tung et al. 2009], it has been shown that the probabilistic fusion of narrow and wide baseline stereo allows us to perform accurate and complete 3D reconstructions of dynamic scenes. Consequently, we propose to take advantage of this theoretical result to derive a novel 3D video reconstruction framework which requires only a minimal set of standard video cameras (cf. Fig. 1). The solution reduces the acquisition cost and reconstruction effort, and should make 3D video systems accessible to a larger public.

2 The minimal concept

As shown in Fig. 1 (Left), almost all regions on a 3D model surface are visible by at least two cameras in a 15 camera setup. Using the classical wide baseline MVS approach, it is impossible to recover surface regions where $N = 1$ (usually concave regions). However, as we are dealing with models in motion, it is possible to apply structure-from-motion (or SfM) techniques to reconstruct surface regions that are visible by only one camera. In particular, the image content stability provided by each single-view video helps to achieve robust narrow baseline stereo matching, thus avoiding the issues encountered due to lighting variations from different distant viewpoints.

3D video setups require several video cameras to maximize regions of overlap because redundancies are necessary to ensure good quality reconstructions. In our proposed framework, information redundancies are not as crucial. For example, a minimal setting requires 6 surrounding XGA video cameras, so that the angle between two cameras (a stereo pair) is 60 degrees, and one additional camera for a view from the top as shown in Fig. 1 (Center). Unfortunately a complete and accurate reconstruction is impossible using SfM only. Hence we propose to compute a probabilistic estimation of the true model surface: 3D structures recovered from motion and sparse 3D features obtained by robust stereo photo-consistency are fused into a maximum a posteriori Markov random field problem which can be solved with global optimization algorithms [Tung et al. 2009]. The posterior probability to maximize is:

$$p(\Theta|\Phi,\Gamma) \propto \prod_{i} E_p(\theta_i,\phi_i)E_q(\theta_i,\gamma_i) \prod_{i} \prod_{j\in\mathcal{N}(i)} V(\theta_i,\theta_j), \quad (1)$$

where $\Theta = \{\theta_i\}$ are 3D surface points at t, $\Phi = \{\phi_i\}$ are input images, $\Gamma = \{\gamma_i\}$ are structures from motion flows, E_p is a local evidence based on the photo-consistency score estimated at Θ from Φ, E_q is a local evidence based on the distance between Γ and sparse 3D features derived from Φ, $N(i)$ represents the neighbors of node i and V is a smoothness assumption. Preliminary experiments have shown promising results on synthetic and real-world datasets. Thus, minimal 3D video is a novel solution to produce accurate 3D videos with a minimal setting.

References

DE AGUIAR, E., STOLL, C., THEOBALT, C., AHMED, N., SEIDEL, H.-P., AND THRUN, S. 2008. Performance capture from sparse multi-view video. *ACM Trans. Graphics 27*, 3.

TUNG, T., NOBUHARA, S., AND MATSUYAMA, T. 2009. Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. *Proc. IEEE Int'l Conf. Computer Vision.*

[*] e-mail:tung@vision.kuee.kyoto-u.ac.jp
[†] e-mail:tm@vision.kuee.kyoto-u.ac.jp