Simultaneous super-resolution and 3D video using graph-cuts

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

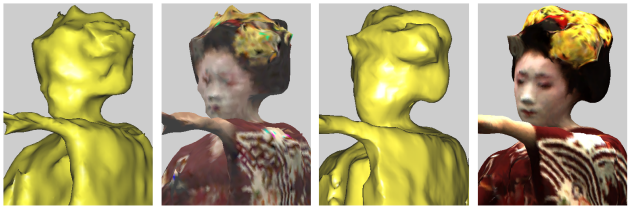
Abstract

This paper presents a new method to increase the quality of 3D video, a new media developed to represent 3D objects in motion. This representation is obtained from multi-view reconstruction techniques that require images recorded simultaneously by several video cameras. All cameras are calibrated and placed around a dedicated studio to fully surround the models. The limited quality and quantity of cameras may produce inaccurate 3D model reconstruction with low quality texture. To overcome this issue, first we propose super-resolution (SR) techniques for 3D video: SR on multi-view images and SR on single-view video frames. Second, we propose to combine both super-resolution and dynamic 3D shape reconstruction problems into a unique Markov Random Field (MRF) energy formulation. The MRF minimization is performed using graph-cuts. Thus, we jointly compute the optimal solution for super-resolved texture and 3D shape model reconstruction. Moreover, we propose a coarse-to-fine strategy to iteratively produce 3D video with increasing quality. Our experiments show the accuracy and robustness of the proposed technique on challenging 3D video sequences.

1. Introduction

3D video is a new media developed these last years to represent 3D objects in motion [13, 17]. This technique captures dynamic events in the real world. It records time varying 3D models with surface properties such as color and texture. Its applications cover wide varieties of personal and social human activities: entertainment, education, sports, medicine, culture, heritage preservation and so on.

Models are recorded by several video cameras in a dedicated studio. Every 3D video frame contains one or several 3D textured meshes, each frame being acquired at video rate. Dynamic 3D model reconstructions are performed using multi-view stereo techniques (one model is computed for each frame). Therefore input image quality is a crucial factor to obtain high-quality (detailed) 3D video. For static 3D model reconstruction, several high-quality methods exist [22] where the image quantity infers directly on the quality of reconstruction. In 3D video, models are in motion. And due to the limited number of cameras, details on object surface cannot always be recovered and mapped textures may lack of precision as well. A current framework uses a deformable mesh model to estimate 3D shape from multi-view videos [17]. The deformation process of the mesh model is heterogeneous and minimizes an energy. Each vertex of the mesh changes its position according to its photometric property (i.e. if it has prominent texture or not), and physical property (i.e. if it is on a rigid part of the object or not). This heterogeneous deformation model enables to reconstruct 3D models with a single and unified computational framework. To increase the 3D video quality, we propose super-resolution (SR) techniques dedicated to multi-view videos. In particular, our approach focuses on both SR on multi-view images and SR on single-view video frames. Problems are formalized as Markov Random Field (MRF) energy models with maximum a posteriori (MAP) estimation. We propose to take advantage of the graph-cuts (min-cut/max-flow theory) to minimize MRF energies [5]. Moreover we combine SR reconstruction with 3D object shape optimization, as it is well known that graph-cuts suit very well to multi-view stereo reconstruction [14, 30]. Therefore the global energy minimization computes simultaneously the optimal 3D model shape and texture (cf. Figure 1). Moreover we propose a strategy to obtain iteratively coarse-to-fine 3D video reconstruc-



(a) 3D video (b) Super-resolution 3D video

Figure 1. **Super-resolved 3D video**. Our super-resolution methods increase the quality of 3D video. Texture resolution is higher, and mesh surface gains geometric details. (a) shows a current frame of 3D video. (b) shows our super-resolution 3D video reconstruction.

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

tion. Our experiments show the accuracy and robustness of the proposed technique on challenging 3D video sequences. For examples, videos of Japanese traditional dancers were used. Gestures are sharp and clothing is very fine. Thus a reliable storage media is necessary.

The rest of the paper is organized as follows. The next section discusses work related to the techniques presented in this paper (3D video, super-resolution and multi-view stereo using graph-cuts). Section 3 presents super-resolution on multi-view images acquired by video cameras. Section 4 presents super-resolution on single-view video frames. Section 5 describes the coarse-to-fine reconstruction strategy and the compound energy to minimize to reconstruct simultaneously high-quality texture and 3D shape of 3D video. Section 6 presents experimental results. Section 7 concludes with a discussion on our contributions.

2. Related work

Since these last years computer hardware devices became powerful enough to handle heavy calculations. Hence, nowadays an increasing number of research groups are involved in dynamic 3D multi-view stereo reconstruction [13, 9, 17, 12, 25]. The main motivation is to model a realistic virtual world. The 3D video sequences shown in this paper were acquired in real-time using a cluster of 15 node PCs and 15 cameras. The reconstruction method relies on a shape-from-silhouette approach to produce rough 3D mesh model sequences. Then 3D mesh surfaces are optimized using a cost function to minimize [17, 19]. As well, texture maps are computed for each video frame. Numerous multi-view stereo reconstruction algorithms can be found in the literature (e.g. see [22] for a recent survey). To obtain very accurate 3D models from stereo techniques, it is well known that lots of high resolution images are required. In the 3D video framework, we cannot afford to set several ten high-resolution cameras or so in a studio. Thus having 15 video cameras means 15 images for each 3D model reconstruction. As shown in Figure 2, a limited number of images makes difficult to recover all 3D shape details. Camera calibration errors are not well compensated with few cameras and pixel photo-consistency is weak. To guarantee good 3D video reconstruction, a hardware-based solution is to use HD video cameras as in [25].

The solution we propose is to increase the quality of 3D video reconstruction by applying super-resolution algorithms dedicated to 3D video. Super-resolution (SR) is a technique to recover detailed information from degraded data. Therefore one can reconstruct high resolution (HR) images from several low resolution (LR) images. It has been extensively used in numerous applications such as photography enlargement, video surveillance, medical imaging, satellite imaging, etc. Lots of previous work can be found in the literature (e.g. see [31] and [4, 7, 20] for reviews). To

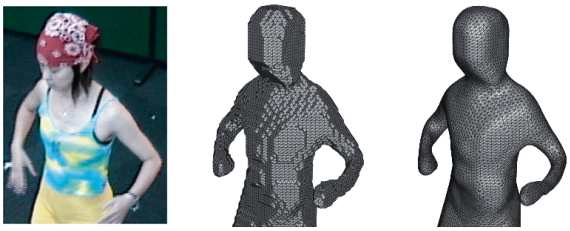


Figure 2. **3D video reconstruction using shape-from-silhouette and deformable 3D mesh model**. (Left) shows an original video frame. (Middle) shows an initial mesh (the visual hull) obtained by shape-from-silhouette technique. (Right) shows a mesh after the deformation step. Due to lack of photo-consistency, high frequency regions are not well reconstructed (e.g. the head).

obtain an HR image from multiple LR images, LR images have to provide different views of the same scene with different subpixel shifts. Usually, the methods use a model of the image formation process to relate N LR images y_k to an HR image x [8]:

$$y_k = DB_kW_kx + \epsilon_k, \qquad 1 \leq k \leq N. \tag{1}$$

This model takes into account optical distortion (warp matrix W_k), the camera point spread function (blur matrix B_k), aliasing (sub-sampling matrix D), and additive noise ϵ_k. It is an ill-posed inverse problem which does not have a straightforward solution and usually requires some additional regularization. However, various approaches have been proposed [34, 1, 11, 23]. In general, SR image reconstruction methods consist in three steps: 1) registration of LR images or motion estimation, 2) interpolation on an HR grid, and 3) restoration for blur and noise removal. Besides, if there is a sufficient number of LR images, regularization methods (e.g. stochastic approaches) can be used to stabilize the inversion of the ill-posed problem. In [21] SR was applied on still images for 3D reconstruction of a book page in a hemispheric configuration with 51 cameras. In our 3D video framework, we address both SR on multi-view images and SR on single-view video frames. As a matter of fact, increasing the quality of images will produce more accurate 3D shapes and better textures. Our proposed SR approach uses Markov Random Field energy formulations that suit dynamic 3D shape reconstruction.

As presented in [26], several problems in vision can be formulated as a Markov Random Field (MRF) energy, and solved using minimization techniques. Energy minimization is a difficult task as it usually requires a lot of computational time to eventually find local or global minima. Classical techniques such as simulated annealing are very slow in practice. Fortunately, graph-cut approaches have been developed in the last few years [6]. The idea is to build a dedicated graph for the energy function to be minimized; then the minimum cut on the graph minimizes the energy as well

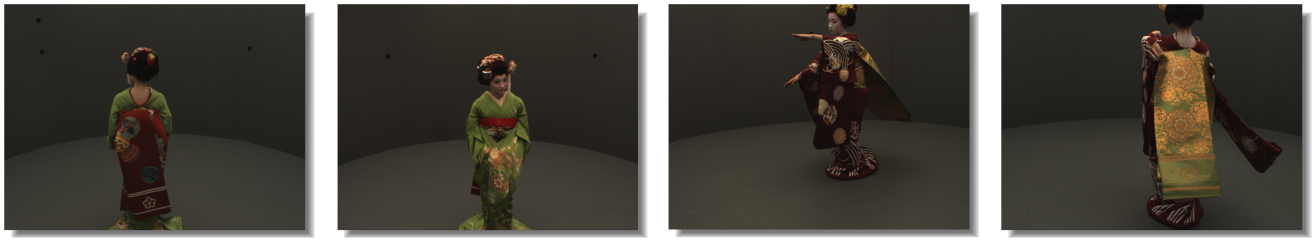


Figure 3. **Different camera views of models**. Each camera produces a video at resolution 1024*768 pixels. High fidelity media is necessary to represent fine details of Japanese traditional clothing. Note that less than one third of the images contains relevant information.

(either globally or locally). Using max flow algorithms, the minimum cut can be computed very efficiently. In general the solution found has interesting theoretical quality guarantee. These methods have been successfully used for a large variety of applications, such as 3D multi-view reconstruction [14, 30] or even super-resolution on images with synthetic noise and small displacements [18].

Therefore we propose to combine both super-resolution and dynamic 3D shape reconstruction problems as an original MRF energy formalization which minimization is performed using graph-cuts [15]. We compute simultaneously the optimal solution for super-resolved texture and dynamic 3D shape model reconstruction. The result is a 3D video with increased quality.

3. Super-resolution on multi-view images

In this section, we present super-resolution on a set of multi-view images acquired by calibrated video cameras. Each camera produces video frames of size $s = 1024 * 768$ pixels (cf. Figure 3). Colors are encoded on 24 bits (RGB channels). Obviously interpolation methods could be applied to obtain SR images. Nevertheless they may lead to the loss of details in high frequency regions. Hence a regularization technique using MAP-MRF formalism was preferred.

The proposed super-resolution approach consists in: 1) magnification of every image from low resolution (LR) to high resolution (HR), and accurate alignments of LR images onto HR grids to gain subpixel information in each HR image; 2) HR image regularization using MRF energy formulation and minimization with graph-cuts.

3.1. Multi-view registration

Assuming N images taken by N calibrated cameras, and a magnifying factor m (e.g. $m = 2$), the image alignment algorithm is the following:

1. Compute a visual hull of the model using silhouette projection intersections as in [17].

2. Consider each image plane of the LR images L_i ($i \in [1, N]$) with the magnification factor m. Thus we obtain HR grids H_i of size $m * m * s$ pixels.

3. Using the prior knowledge on the model 3D shape (i.e. the visual hull from step 1) and Z-buffer computations, project every visible pixel from LR images L_j ($j \in [1, N], j \neq i$) onto the H_i grids. Thus we obtain HR images of non-occluded regions with subpixel precision (cf. Figure 4). Artifacts due to calibration errors are compensated using consistency criterion on pixel colors [3].

4. For every pixel p of the HR grid, if no visible pixels from any LR images were projected on p, then set the value of p as the interpolation of neighborhood values (e.g. using a 3*3 mask).

Finally we obtain a set of N HR images $m * m$ times bigger than the LR images, including more accurate information. Nevertheless, a regularization step is necessary to overcome remaining artifacts and give a smooth solution.

Figure 4. **Projection of all visible pixels from low resolution images on a high resolution (HR) grid**. (Left) Original frame. (Right) Blank HR grid with subpixel information. Inconsistent pixels were filtered.

3.2. MRF energy formulation

In our framework, LR images are acquired by calibrated video cameras. Blur effects are limited and do not necessarily need to be simulated as in [18] (where SR is applied on images with synthetic noise and small displacements). Moreover we compute optimization with respect to *observed* HR grids as previously defined, instead of considering every projected pixel from LR images. The MRF en-

ergy is formulated to enable discontinuity preservation and image regularization using graph-cuts.

Let $\mathcal{P} = \bigcup_{i=1}^{N} \mathcal{P}_i = \{p\}$ be the set of pixels of HR images H_i, and $\mathcal{L} = \{l_p\}$ be a discrete set of labels corresponding to the possible pixel values of $\mathcal{P}$. HR images are optimized using MRF energy formulation:

$$E(f) = E_d(f) + E_s(f). \tag{2}$$

$E(f)$ is the energy of the labeling $f : \mathcal{P} \rightarrow \mathcal{L}$. $E_d(f)$ is the data energy measuring the disagreement between f and $\{p\}$:

$$E_d(f) = \sum_{p \in \mathcal{P}} D_p(l_p) = \sum_{p \in \mathcal{P}} (l_p - i_p)^2, \tag{3}$$

where $D_p(l_p)$ stands for the cost of a label l_p on pixel p, i_p being the intensity of p. $E_s(f)$ is the neighbor smoothness cost:

$$E_s(f) = \sum_{\{p,q\} \in \mathcal{N}} V(p,q) = \sum_{\{p,q\} \in \mathcal{N}} \lambda \cdot \min(K, |l_p - l_q|), \tag{4}$$

where $\mathcal{N}$ is the set of neighborhood configurations in a 4-connected neighborhood system, and $V(p,q)$ is the non-truncated linear cost for constant $\lambda = 2$ and $K = 255$.

Then $E(f)$ is minimized using graph-cuts with an expansion algorithm as described in [6]. Note that one energy is formulated for each R, G, B channel. An example of results is presented in Figure 5.

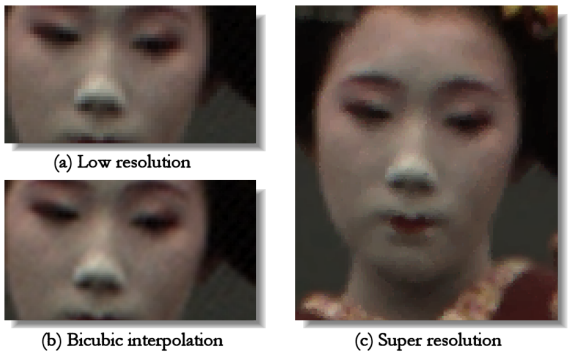


Figure 5. **Multi-view image super-resolution reconstruction**. (a) is the original frame: obviously aliasing effects are visible. (b) has been obtained with bicubic interpolation of the LR image: as expected, the result is smooth and we tend to lose high frequencies. (c) is the super-resolved image: discontinuities are well preserved and video artifacts removed.

4. Super-resolution on single-view video frames

In this section, we consider each video camera independently. Consecutive frames from a single-view video sequence contain a lot of redundancies. Thus it is possible to increase the quality of each camera video sequence. As in Section 3, the super-resolution reconstruction approach consists on LR image registration to gain subpixel details in overlapping regions, and optimization of HD grid using graph-cuts. Indeed, the quality of super-resolved images highly relies on the correctness of image alignments between consecutive frames (e.g. see [33, 32] for surveys on image registration). In [2], a piecewise image registration method was proposed for large motions. It relies on a multi-label graph-cut optimization to estimate dense motion field between two images. Here, we address the problem of single-view multiple frame registrations with no necessary large motions. We employ robust feature detection and matching to accurately estimate motion between the image pairs. An efficient process is used to detect and discard incorrect matchings which may degrade the output quality. As similar regions of interest (ROI) are detected in two consecutive frames L_t and L_{t+1}, we propose to use a local mapping model [10] to transform the ROI of L_t onto L_{t+1} as a warped texture. The ROI are formed by the pixels included in the triangulated area of the detected features.

4.1. Feature matching

Automatic image alignments can be performed either by pixel-based alignments (e.g. optical flow estimation [24]), or by feature-based alignments (motion estimation). In our framework, we are interested by tracking regions of interest in consecutive frames (regions on the model). The Scale Invariant Feature Transforms (SIFT) detector by [16] is efficient and reliable to our purpose. It is invariant to rotation and scale, robust to change in lighting, and encodes local area brightness patterns. We use the SIFT feature matching method proposed by [16] as well (cf. Figure 6). A post-processing step based on the SIFT feature vectors removes the outliers from the motion field.

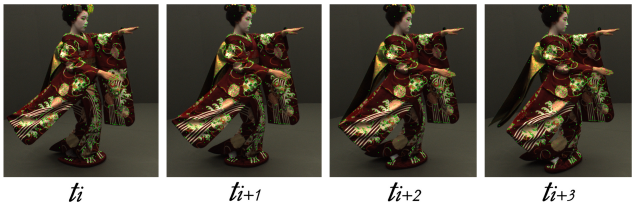


Figure 6. **SIFT feature detection in consecutive frames**. The kimono has a lot of details. Close to 1400 SIFT features were found on each of these images.

4.2. Texture warping and mapping

We propose to create a mesh based on the detected features. The ROI is then the mesh texture. Considering two consecutive frames L_t and L_{t+1}, Delaunay triangulation is performed on L_t and projected onto the detected features in

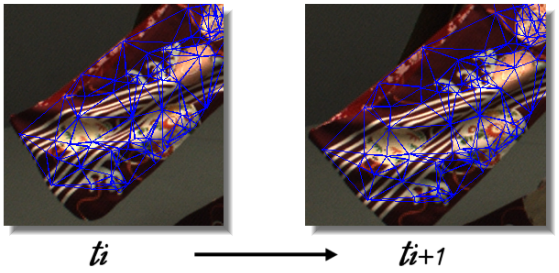


Figure 7. **Local mapping model**. Similar region of interests (ROI) are tracked along the sequence. Here, triangulations of ROI are similar in consecutive frames. Then textures are extracted and mapped onto an HD grid using a piecewise cubic mapping.

L_{t+1}. If the features are well detected in both frames L_t and L_{t+1} and matched, then meshes have same topology in each frame. Obviously, unfiltered outliers from the previous step would return wrong triangulation in L_{t+1}. Nevertheless in consecutive images of a 3D video sequence similar features and outliers are well detected. Finally, we extract the texture from the mesh of L_t and apply it onto the mesh of L_{t+1} using a piecewise cubic mapping function [10]. Indeed, textures of similar ROI are wrapped onto an HR grid. If the ROI area in the image L_t were bigger than the ROI area in the image L_{t+1}, then the mapping of the texture from L_t to L_{t+1} would give more details to L_{t+1}. This would lead to an HR image with subpixel informations (cf. Figure 7).

The texture mapping is applied on every pair of consecutive frames having the similar ROI. ROI are tracked using the feature matching process described in the previous subsection. Afterward, as all similar ROI are extracted, textures are mapped onto ROI HR grids. Pixels that are not color-consistent are filtered using a dissimilarity measure [3]. Then we obtain compound HR images having subpixel details from LR frames. Finally, we use an MRF energy formulation as in Section 3 to regularize HR images and reconstruct SR images. Minimization is performed using graph-cuts (cf. Figure 8).

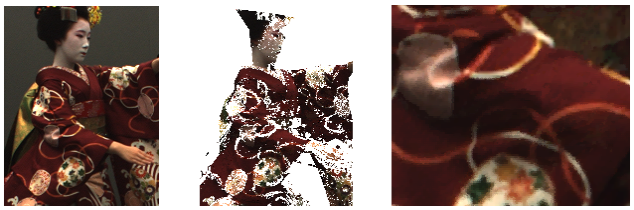


Figure 8. **Super-resolution reconstruction of single-view video sequence**. (Left) shows an original frame. (Middle) shows the projection of the relevant pixels from 10 consecutive frames onto a blank HR grid. (Right) shows a detail of the super-resolved reconstructed image.

5. Simultaneous super-resolution and 3D video

We propose a coarse-to-fine approach to iteratively reconstruct increasing quality 3D video. As presented in previous sections, our super-resolution and 3D video approaches use MRF energy formalism. Therefore we propose to combine both problems in one unique energy. The minimization of this new global energy returns simultaneously the optimal 3D video reconstruction with the optimal texture. Fortunately this can be achieved efficiently using graph-cuts [6, 15].

5.1. Coarse-to-fine strategy

Our coarse-to-fine scheme is illustrated in Figure 9. It produces a 3D video whose quality increases iteratively as videos are acquired. We assume video frames are acquired simultaneously at every iterative step at time $t_i, i \geq 0$. Starting at t_0, SR reconstruction is applied on the set of multi-view LR images (cf. Section 3) acquired at each iteration t_i. Meanwhile SR is applied on every single-view video frame (cf. Section 4) after each iteration. The SR reconstruction involves all previously reconstructed super-resolved frames at $t < t_i$, and the set of super-resolved images reconstructed from the set of LR images at t_i. At each iteration step t_i, SR reconstruction is computed simultaneously with 3D model shape using the MRF energy minimization with graph-cuts. Thus a super-resolved 3D video frame can be reconstructed at each iterative step.

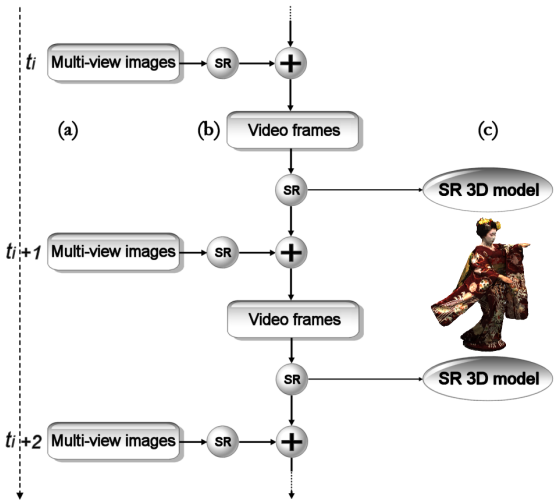


Figure 9. **Iterative coarse-to-fine strategy.** a) Super-resolution on multi-view images at time $t = t_i$: image alignments and regularization. b) Super-resolution on every single-view video frame for $t \leq t_i$: feature detection, region of interest tracking, texture projection and regularization. c) Simultaneous super-resolution and 3D video reconstruction using a combination of MRF energy formulation. Minimization of the compound energy is performed using graph-cuts.

5.2. Simultaneous energy minimization

Assuming g is a 3D position labeling, and f is the color labeling as described in Section 3.2, the global compound energy $\tilde{E}(g)$ formulation for simultaneous SR and 3D shape reconstruction using graph-cuts can be written as follows:

$$\tilde{E}(g) = E(f(g), g), \quad (5)$$

$$f(g) = \arg\min_{f \in \mathcal{P}} E(f, g), \quad (6)$$

$$E(f, g) = E^{SR}(f) + E^{3D}(g), \quad (7)$$

where $E^{SR}(f)$ is the super-resolution energy of labeling f presented in Section 3.2, and $E^{3D}(g)$ is the energy of labeling g for multi-video 3D reconstruction. We have revisited the energy $E^{3D}(g)$ with silhouette constraints from [17] to handle volumetric graph-cut framework in an MRF formalism as [25, 30]. To impose silhouette constraints, contour generators (CGs) are first explicitly estimated by dynamic programming. Then estimated CGs are used as *fixed* voxels in the graph structure. Our approach is somehow similar to the one presented in [27], but we use dynamic programming to estimate the optimal CGs considering continuity (smoothness) between estimated points instead of computing each point individually.

Suppose we have N cameras and a binary silhouette image S_i for each camera C_i ($i \in [1, N]$). We assume the outlines of S_i consist in closed curves with ring topology. We denote the j-th outline by $s_{i,j}$, and the x-th pixel of $s_{i,j}$ by $s_{i,j}(x)$ ($x \in \left[1, N_{s_{i,j}}\right]$) where $N_{s_{i,j}}$ is the number of pixels of $s_{i,j}$. Every 2D point of silhouette outline $s_{i,j}(x)$ has one or more corresponding 3D points $\{P_{i,j}(x)\}$ lying on the object surface. We can expect that each has high photo-consistency with camera image pixels. The object visual hull gives the possible 3D positions of CGs for each $s_{i,j}(x)$. We formalize the contour generator estimation problem as an energy minimization problem of a function E_{cg} defined as follows:
$E_{cg} = \sum_i E_{S_i}$, $E_{S_i} = \sum_j E_{s_{i,j}}$, $E_{s_{i,j}} = \sum_x E(x)$, and

$$\sum_x E(x) = \sum_x E_p(p_x) + \lambda \sum_x E_d(p_x - p_{x+1}), \quad (8)$$

where p_x denotes the selected 3D point from $P_{i,j}(x)$ corresponding to $s_{i,j}(x)$, $E_p(p_x)$ is the photo-consistency term at p_x, and $E_d(p_x - p_{x+1})$ is the distance between p_x and p_{x+1} as a smoothness term. We use the dissimilarity function of [3]. Let C_{p_x} denote the set of cameras which can observe a 3D point p_x on V. We define our photo-consistency function as:

$$E_p(p_x) = \sum_{c_i, c_j \in C_{p_x}} \exp(e_i.e_j - 1) \cdot d(p_x^{c_i}, p_x^{c_j}), \quad (9)$$

where c_i and c_j denote a pair of cameras in C_{p_x}, e_i and e_j are the viewing direction of c_i and c_j respectively, $\exp(e_i \cdot e_j - 1)$ is a scalar weight based on the camera directions, $p_x^{c_i}$ and $p_x^{c_j}$ denote 2D projections of p_x at c_i and c_j respectively. The function $d(p_x^{c_i}, p_x^{c_j})$ is the pixel-based dissimilarity function defined by [3]. Since we assumed that $s_{i,j}$ is a closed curve and parameterized by a single variable x, the smoothness term of x depends only on its neighbor $x + 1$. Hence this minimization problem can be solved efficiently by dynamic programming whereas a min-cut problem framework would be computationally expensive. We denote V_{cg} as the optimal set $\{p_x\}$ minimizing E_{cg}.

As proposed in [30], we use the visual hull V as the initial estimation of the object shape, and use graph-cuts to compute the optimal surface S^* minimizing the photo-consistency on S^* under contour generator constraints. We define the graph $\mathcal{G}$ so that the min-cut returns the object shape on its source side. Each voxel $v \in V$ is associated to a node of $\mathcal{G}$ connected by edges in a 6-neighborhood system. The cost of an edge between two voxels v_i and v_j is $W_{ij} = E_p(\frac{p_i + p_j}{2})$, where p_i and p_j are the 3D positions of v_i and v_j respectively. We assume that the visibility of a point $p \in V$ is approximated by the visibility of its closest voxel lying on the surface of V. Each voxel $v \in V$ is also connected to the source with a constant ballooning cost W_b and to the sink with no cost. Thus, in addition to the graph structure, we set the cost to the source to infinite for every voxel $v_{cg} \in V_{cg}$ so that the minimum cut includes voxels of V_{cg} (cf. Figure 5.2).

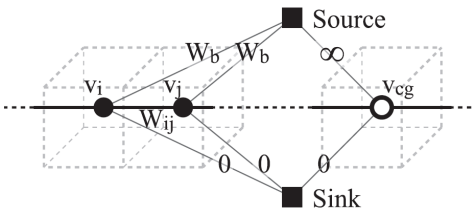


Figure 10. **Graph structure for volumetric graph-cut minimization.** Each voxel is connected to its neighbors with a weight W_{ij}, and to the source with a weight W_b. Each v_{cg} represents a silhouette constraint. It is connected to the source with an infinite weight. All voxels are connected to the sink with a null weigth.

6. Results

The studio has the following configuration: diameter is 6 m, height is 2.5 m, and the size of the area where an object can be reconstructed without defect is approximately 3*3*2 m^3 in the center of the studio. Models are captured using a PC cluster system composed of 15 PCs and one master PC. Every PC has one fixed camera, and cameras are connected to an external pulse generator for triggering. PC specifications are Pentium III 1 GHz * 2, RAM 1 GB. Cameras are Sony XCD-X710CR, XGA. This system delivers synchronized multi-viewpoint images at 25 fps.

(a) 3D video reconstruction

(b) Super-resolution 3D video reconstruction

Figure 11. **Super-resolution of 3D video**. Using our 3D video super-resolution technique, surface and texture of 3D models have better quality and look more realistic. (a) shows a shaded mesh and textured meshes obtained by 3D video reconstruction as in [17], and (b) shows our results with super-resolved images. Details on the kimono are better rendered.

For experimental tests we have worked with video sequences of Japanese dance performed by maikos (apprentice geishas). The elegant appearance of maiko includes colorful clothing, hairstyle, accessories, and make-up. In particular a maiko wears a kimono (traditional Japanese robe style) with an obi (kimono waist band). Obtaining a nice 3D model reconstruction is therefore challenging as cloths are large, non-rigid and contain lots of details to recover. In fact, the video cameras used to capture the relatively fast dances have very short shutter speed (1 ms for this sequence). Hence the image quality is quite low compared to images of a static object captured by still cameras or movie cameras with long shutter. Moreover, acquiring a static object with a single camera allows getting homogeneous image properties from every viewpoints such as gain, color-balance, noise-level, etc. Since we deal with dynamic objects we have to cope with different cameras at different viewpoints, with different properties.

Although our 3D video super-resolution reconstruction requires several steps, the whole pipeline is fully automatized. Super-resolution calculations (cf. Section 3 and 4) were performed on a laptop with Pentium M processor 1.60 GHz and RAM 512 MB. The most expensive step is the SR energy minimization. It is performed for each color channel with 256 labels. One SR computation on HR grid channel requires 4 min. The energy minimization method has been tested against graph-cuts with alpha-expansion and swap algorithm, iterated conditional modes (ICM), and max-product loopy belief propagation (LBP) [6]. For our application, graph-cuts with alpha-expansion algorithm offers the best trade-off between speed and reconstruction quality.

The current implementation takes 5 min to generate a final 3D mesh from multi-viewpoint foreground and silhouette images (Core2 processor 2.4 GHz). This process includes: 1) visual hull reconstruction, 2) contour generator estimation by DP, 3) photo-consistency computation, 4) optimal surface extraction by graph-cut, and 5) per-vertex coloring. The coarse-to-fine strategy allows increasing the quality of 3D video by reconstructing SR images at each iteration step (cf. Section 5.1). In fact, due to hardware limitation it is difficult to increase the size of HR grids at every iteration as magnifying four times initial LR images is already highly memory consuming. However the coarse-to-fine scheme can be used to combine the both presented SR methods and produce SR 3D video without necessarily increasing the HR grid size at each iteration step. According to our experiments, even a magnification factor of 2 is enough to produce fine reconstructions. Figures 1 and 11 illustrate results of our proposed SR 3D video technique: SR allows recovering much more details than current 3D video reconstruction method [17]. Discontinuities are better preserved. In Figure 1, we can observe that the head shape is better recovered with SR. Moreover, the use of SR images has improved significantly the texture quality. The clothing is far better rendered and more colorful, as video noises were handled (regularized). We have obtained reconstructions with voxel size up to 1 mm. To appreciate the sharpness of our results, SR models on Figure 11 can be compared to images on Figure 3 as well.

7. Conclusion

3D video is a new media developed these last years to represent 3D objects in motion[13, 17]. We present an original method to increase the quality of 3D video. We propose super-resolution methods dedicated to 3D video: on multi-view images and single-view video sequences. In particular, we use a practical MRF energy formulation that can be efficiently minimized with graph-cuts. In addition, a compound energy formulation combines super-resolution reconstruction and dynamic 3D shape reconstruction. Hence the energy minimization simultaneously optimizes super-resolution of model texture and 3D shape. Moreover a coarse-to-fine strategy is proposed to iteratively increase the 3D video quality at each step of 3D model reconstruction. Our experiments show the effectiveness of the approach. Hence the quality of 3D videos acquired with low resolution videos can be enhanced.

For further work, to eventually limit the storage cost of

super-resolved texture maps, our approach may be combined to a 3D video compression technique as recently presented in [29]. It takes advantage of the augmented multiresolution Reeb graph[28] properties to store the relevant information of 3D models such as shape, topology and texture. In particular, only one texture map is encoded for a whole 3D video sequence. Therefore, we could tune our coarse-to-fine strategy to update iteratively the texture map, and finally obtain a compact super-resolved 3D video.

Acknowledgement

This research was supported in part by MEXT of Japan under the GCOE program "Informatics Education and Research Center for Knowledge-Circulating Society" and "Development of High Fidelity Digitization Software for Large-Scale and Intangible Cultural Assets" project. In addition, the authors would like to thank Pr. Francis Schmitt from TELECOM ParisTech for helpful comments. Special thanks goes to Ms. Guech Seam Thang for fruitful advice and faithful support.

References

[1] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *TPAMI*, 24(9):1167–1183, 2002.

[2] P. Bhat, K. C. Zheng, N. Snavely, A. Agarwala, M. Agrawala, M. F. Cohen, and B. Curless. Piecewise image registration in the presence of multiple large motions. *CVPR*, pages 2491–2497, 2006.

[3] S. Birchfield and C. Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *TPAMI*, 20(4):401–406, 1998.

[4] S. Borman and R. Stevenson. Spatial resolution enhancement of low-resolution image sequences a comprehensive review with directions for future research. *tech. rep., LISA, Univ. of Notre Dame, Notre Dame, Ind, USA*, 1998.

[5] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *TPAMI*, 26(9):1124–1137, 2004.

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

[7] D. Capel and A. Zisserman. Computer vision applied to super resolution. *IEEE SPM*, 20(3):75–86, 2003.

[8] M. Elad and A. Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE Trans. on Image Processing*, 6(12):1646–1658, 1997.

[9] J. S. Franco, C. Menier, E. Boyer, and B. Raffin. A distributed approach for real-time 3d modeling. *CVPR Workshop*, page 31, 2004.

[10] A. Goshtasby. Image registration by local approximation methods. *Image and Vision Computing*, 6(4):255–261, 1988.

[11] Z. Jiang, T. Wong, and H. Bao. Practical super-resolution from dynamic video sequences. *CVPR*, pages 16–22, 2003.

[12] T. Kanade and P. J. Narayanan. Historical perspectives on 4d virtualized reality. *CVPR Workshop*, page 165, 2006.

[13] T. Kanade, P. Rander, and P. J. Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE Multimedia*, pages 34–47, 1997.

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

[15] V. Kolmogorov and R. Zabih. What energy functions can be minimized via via graph cuts? *TPAMI*, 26(2):147–159, 2004.

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

[17] 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.

[18] U. Mudenagudi, R. Singla, P. K. Kalra, and S.Banerjee. Super resolution using graph-cut. *ACCV*, pages 385–394, 2006.

[19] S. Nobuhara and T. Matsuyama. Deformable mesh model for complex multi-object 3d motion estimation from multi-viewpoint video. *3DPVT*, pages 264–271, 2006.

[20] S. C. Park, M. K. Park, and M. G. Kang. Super-resolution image reconstruction: A technical overview. *IEEE SPM*, 20(3):21–36, 2003.

[21] H. Saito. Super resolving texture mapping from multiple view images for 3d model construction. *IEEE Int. Conf. on Systems, Man, and Cybernetics*, 2:1418–1421, 2000.

[22] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. *CVPR*, pages 519–526, 2006.

[23] E. Shechtman, Y. Caspi, and M. Irani. Space-time resolution in video. *TPAMI*, 27(4):531–545, 2005.

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

[25] J. Starck and A. Hilton. Surface capture for performance based animation. *IEEE CGA*, 27(3):21–31, 2007.

[26] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields. *TPAMI, to appear*.

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

[28] T. Tung and F. Schmitt. The augmented multiresolution reeb graph approach for content-based retrieval of 3d shapes. *Int. Jour. of Shape Modeling*, 11(1):91–120, 2005.

[29] T. Tung, F. Schmitt, and T.Matsuyama. Topology matching for 3d video compression. *CVPR*, 2007.

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

[31] J. Yu, L. McMillan, and S. Gortler. Scam light field rendering. *Pacific Graphics*, pages 137–144, 2002.

[32] L. Zagorchev and A. Goshtasby. A comparative study of transformation functions for nonrigid image registration. *IEEE Trans. on Image Processing*, 15(3):529–538, 2006.

[33] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21:977–1000, 2003.

[34] A. Zomet, A. Rav-Acha, and S. Peleg. Robust super-resolution. *CVPR*, pages 645–650, 2001.