Estimating a Patient Surface Model for Optimizing the Medical Scanning Workflow

Vivek Singh[1], Yao-jen Chang[1], Kai Ma[1], Michael Wels[2], Grzegorz Soza[2], and Terrence Chen[1]

[1] Imaging and Computer Vision
Siemens Corporation, Corporate Technology, Princeton, NJ, USA
[2] Siemens AG, Healthcare Sector
Forchheim, Germany

Abstract. In this paper, we present the idea of equipping a tomographic medical scanner with a range imaging device (e.g. a 3D camera) to improve the current scanning workflow. A novel technical approach is proposed to robustly estimate patient surface geometry by a single snapshot from the camera. Leveraging the information of the patient surface geometry can provide significant clinical benefits, including automation of the scan, motion compensation for better image quality, sanity check of patient movement, augmented reality for guidance, patient specific dose optimization, and more. Our approach overcomes the technical difficulties resulting from suboptimal camera placement due to practical considerations. Experimental results on more than 30 patients from a real CT scanner demonstrate the robustness of our approach.

1 Introduction

State-of-the-art medical imaging technologies such as CT, MR or PET provide high quality visual images of the inside of the patient body to the radiologists and physicians for better diagnosis. Nevertheless, the workflow of the existing scanning procedure depends heavily on the experience and subjective decisions of the technician, which often results in suboptimal image quality, large inter-technician variability, unnecessary radiation to the patient (in case of CT), and prolonged scanning time. In this paper, we propose to equip the scanner with visual capability and the knowledge about the patient surface geometry to improve scanning in all these aspects.

In order to provide the knowledge about the 3D patient surface, we propose a novel framework to obtain a detailed body surface model of the patient (on the table) using a range imaging device, which can ease scan planning in several ways. The estimated surface model includes a detailed body surface mesh as well as the location of various anatomical landmarks (such as the shoulders, thyroid, etc.) in the coordinate reference frame of the scanner. These surface landmarks provide a rough estimate of the organ positions which enables automatic table height adjustment. Furthermore, it can also be used to restrict the scan range to reduce unnecessary irradiation. Other potential clinical benefits include but are

P. Golland et al. (Eds.): MICCAI 2014, Part I, LNCS 8673, pp. 472–479, 2014.

not limited to motion compensation for better image quality, consistent imaging quality across technicians, automatic sanity check based on movement of patient during scan, optimized dose based on patient body size, breathing motion detection, and much more.

Besides the algorithm for surface estimation, appropriate placement of the range imaging device also plays an important role for the optimal surface estimation. Keeping track of key factors such as patient visibility in the camera's field of view, ease of the installation, cost and sensor noise characteristics, we mounted an ASUS Xtion (a structured light sensor that captures depth and color information) on top of the gantry as shown in Figure 1. The camera was positioned carefully to keep as much of the patient in the view as possible while avoiding occlusion by the gantry surface. Mounting the sensor on the gantry also simplifies the installation procedure and avoids structural modifications to the scanning room. Note that range imaging devices for patient positioning in a medical scanning setup have been discussed in the literature before, although primarily in the context of fractionated Radiation Therapy. For instance, [2] use a Microsoft Kinect sensor for coarse patient setup by aligning the range image data of the patient (on the table) with a previously obtained CT scan. [4] use the laser surface scanning system GALAXY to perform a similar task. In our work, we produce a 3D patient deformable mesh from one single camera shot without prior scan. The information can be used in real time for improving various aspects of subsequent scanning.

We validate the performance of our approach by data captured from 33 people with different body shapes, sizes, clothing and ethnicity. Our results demonstrate that the patient surface can be estimated with high accuracy and can potentially enable aforementioned applications.

Fig. 1. Structured Light Depth Sensor mounted on top of a CT Scanner

2 Machine Learnt Patient Surface Model Estimation

In this section, we describe the approach to estimate the human body surface geometry from data captured from a single snapshot from a range imaging device

such as ASUS Xtion mounted on top of the scanner gantry. While the problem of estimating human body pose and surface geometry in an unconstrained setting is very difficult, we systematically consider the constraints imposed in a medical scanning workflow to solve this problem to a high accuracy. For instance, the knowledge that the patient is lying down on the examination table significantly reduces the degree of articulation of the body pose as well as simplifies the problem of separating the patient region from the background scene information. Furthermore for medical scanning the patient is often required to be in one of the few poses depending on which body region needs to be scanned. For instance, when a head scan is requested, patient pose should be such that the head is close to the gantry and arms are either folded on the abdomen or on the side. While such constraints on body pose help reduce the search space, the system still has to deal with significant shape variations from patient clothing as well as body shape. Furthermore, the depth sensor on the gantry captures data at an angle (about 45 degrees) which can lead to occlusion when a knee rest is used. The approach must also deal with the noise in the data captured from the sensor such as noise due to stereo analysis as well as depth quantization.

In our approach, we systematically take the prior knowledge into account and use machine learning to estimate the patient surface geometry with a high accuracy, while being fast enough for a seamless integration in the patient scanning workflow. Our algorithm consists of 4 modules, as shown in Figure 2 - Patient Isolation (to extract patient region from the rest of the scene), Body Pose Classification (to classify the body pose as prone or supine and head first or feet first), Patient Surface Model Estimation (to fit a kinematic model of body landmarks) and finally, Dense Body Shape Estimation (to fit a deformable body mesh to the range data).

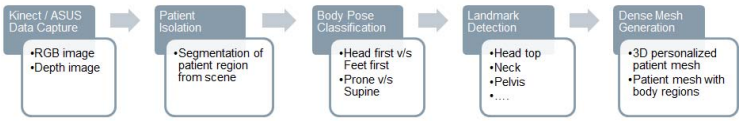

Fig. 2. Processing Pipeline for Patient Surface Model Estimation

2.1 Patient Isolation

Given the color and range data from the sensor (represented as a point cloud), we first localize the image region only containing the patient and the table. Note that the relative position of the range sensor w.r.t. the scanner is known (established during the calibration process) and the range of table movement is limited. We use this spatial prior to automatically crop the image region enclosed by the 3D volume containing the patient and the table. We then transform the cropped data such that the x-axis is aligned with the ground normal and z-axis is aligned with the table length. The transformed depth data (and associated color

information) is then orthographically projected on the y-z plane to generate a color and depth image pair (referred as reprojected image), which is then used for subsequent processing. Next, to further refine the position and extent of the patient, we apply a full body detector on the reprojected image. For detection, we employ Probabilistic Boosting Tree (PBT) [8] with 2D HAAR features extracted over reprojected depth and surface normal data (computed using PCL [5]).

2.2 Body Pose Classification

Given the coarse patient position information, we then classify the patient pose in prone or supine, and head first or feet first. We again employ the Probabilistic Boosting Tree (PBT) [8] for this task; for better results, we extend the PBT framework to multiple channels by considering HAAR features extracted from reprojected depth image, surface normal data, saturation image as well as U and V channels from LUV space. In our experiments, we observed that use of multiple channels provides a significant improvement over only using the depth information.

Instead of training one multi-class classifier, we train multiple binary classifiers to systematically handle the variations in the data. We first apply a head first vs. feet first classifier by considering half of the patient region that is close to the sensor (this region covers the upper half of the body for the head first case and lower half for the feet first case). Based on whether pose is head first and feet first, we then apply a prone vs supine classifier. Note that we train separate prone classifiers based on whether the patient is head first or feet first. This is because when the patient is lying on the examination table the data statistics on the head in the head first case are significantly different compared to the feet first case; this is due to the large angle between the camera and body surface as well as the increase in data noise with increasing distance from the sensor.

2.3 Patient Surface Model Estimation

Given the patient pose information, we then fit a sparse body surface model with anatomical landmarks to the data. The body model is represented as a Directed Acyclic Graph (DAG) over the anatomical landmarks on the body surface, such as thyroid etc., where the graph captures the relative position of the landmarks w.r.t. each other. The patient surface is modeled using 10 body landmarks - head, groin, left and right landmarks for shoulders, waist, knees as well as ankles. For each landmark, we train a multi-channel PBT classifier with HAAR features over the same channels as used to train the pose classifier. Due to the camera and body surface angle as well as sensor noise, the image statistics vary significantly over the body surface. The data distribution over a landmark in the head first case is different from that in the feet first case. Thus, we train separate landmark detectors for both these cases. Note that during inference, since the pose category is already known at this stage, only one set of landmark detectors is applied. The relative position of the landmarks is modeled using a

Gaussian whose parameters are obtained from annotations over a training data set.

During inference, the landmark detectors are applied sequentially taking contextual constraints of the neighboring landmarks into account. For each landmark l_i, we obtain the position hypotheses based on the detector response as well as the position hypotheses for the parent landmarks in the DAG.

$$p(l_i|I, l_j) = \prod p(I|l_i)p(l_i|l_j) \quad where, \;\; l_j \in Parent(l_i) \tag{1}$$

Given the position information for the patient, first the groin landmark detection is applied in the center region. Next the knee detector is applied on the image region estimated based on the constraints from the pose information as well as relative position information from the hypotheses from the groin region. One by one the landmark hypotheses are obtained for each landmark by traversing the DAG. After all the landmark hypotheses are obtained a global reasoning is performed on these hypotheses, to obtain the set of the landmarks with highest joint likelihood based on detection as well as contextual information. Note that this sequential process also handles the size and scale variations across patients of different age, which may be difficult using body pose estimation approaches that first performs a pixel level body region labeling [6]. Furthermore, due to a large angle between camera and body surface, approaches based on body part detection [7] also may not work well due to significant foreshortening of part geometry.

Figure 3 shows landmark detection results on a few images with patients in different poses. For clarity, the figure also shows the body regions estimated by averaging appropriate landmark positions.

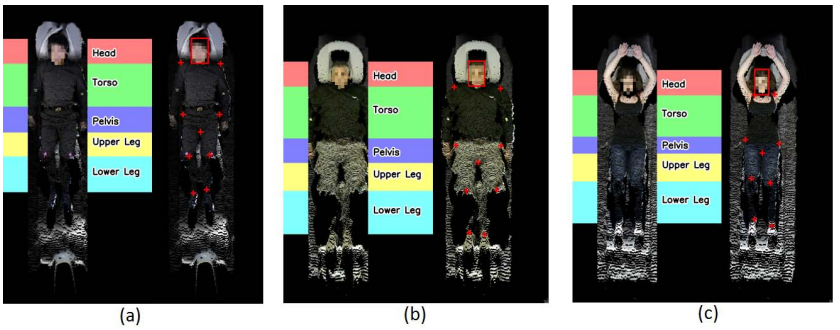

Fig. 3. Landmark detection results and estimated body regions for various poses. Results are rendered on the reprojected images generated by orthographic projection, based on the 3D range data (with associated color information) on the patient table.

2.4 Detailed Body Shape Estimation

Given the location of the landmarks, next we reconstruct the 3D dense patient body surface, represented as a polygon mesh. The reconstructed 3D model is obtained using a parametrized deformable mesh (SCAPE [1]) which can be efficiently perturbed to a target body pose and shape. To model the complex body shape perturbations in compact fashion, [1] decouples the pose and shape perturbation model and during inference, optimizes the pose and shape parameters in an iterative framework. The optimization function for fitting the deformable mesh is modeled as a weighted combination of the surface regularization term (that captures the consistency and smoothness between the neighboring triangles) and a data term (that penalizes for mismatch between the mesh and the input surface data).

In this work, we perform the mesh fitting in a coarse to fine manner for efficiency. We first fit the deformable mesh based on the estimated landmark positions on the patient surface (coarse fit). Next, we apply ICP-based [3] registration to create the correspondences between the deformed template mesh and the 3D surface data. Based on the registration, we optimize the mesh fitting (in an iterative framework) to get a more accurate reconstruction. In our experiments, 2 to 3 iterations were sufficient. Figure 4(a) shows an example of the deformable mesh before and after the optimization. Please observe that the optimized mesh fits the person shape more accurately. Furthermore, the detailed mesh fitting also makes the landmark estimation more precise (as in Figure 4(b)).

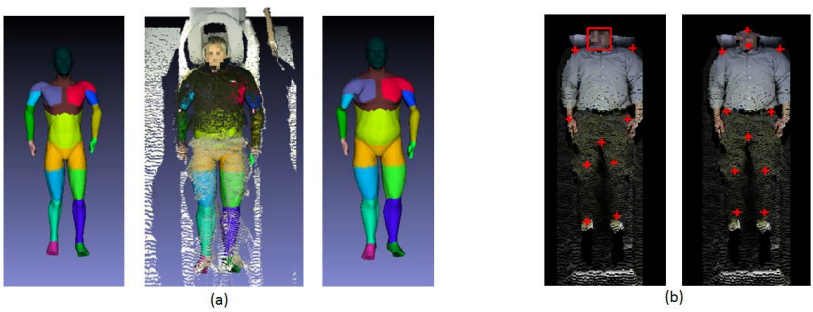

Fig. 4. Detailed Body Shape Estimation. (a) shows the improvement in the body shape estimate before and after the optimization. (b) shows the improvement in landmark precision

3 Experiments

To validate the performance of our approach, we collected data using an ASUS Xtion sensor mounted on the gantry of a Siemens' Somatom CT scanner. The sensor was calibrated w.r.t. the CT scanner by using standard calibration techniques for cameras with color and depth sensors. For our experiments, data was

collected from over 33 people with different age, body shapes/sizes, clothing and ethnicity. For each person, about 40 images were obtained to capture data at different table height and for various body poses that are commonly used in the typical scanning workflow (head first vs feet first, prone vs supine and various hand positions). More than 1000 images were obtained for evaluation purpose.

Although the data was captured at the CT scanner, the people who volunteered for the study were not real patients; hence, we only have access to the color and depth data but not the medical data such as CT or MR. For the purpose of evaluation, we generated ground truth by manual annotation of the anatomical landmarks on the image and depth data. However accurately annotating such landmarks on images can be very difficult and ambiguous due to clothing. For this reason, during the data acquisition process, we captured 2 images for each pose of each patient - one with colored markers at the landmarks and one without any markers. The images with colored markers are only used to assist annotations but not included in the evaluation.

For experimental validation, we split the data into train and test sets of 17 and 16 patients respectively. The total processing time for an image was about 3.5 seconds. We report results for various modules in our processing pipeline. The body pose classification module achieved an accuracy of 0.994178 and 0.989811 for head-first vs feet-first and prone vs. supine, respectively.

To measure the landmark estimation accuracy, we report the error as the distance between the ground truth location and the estimated location along the z-axis. This measure is suitable to evaluate the accuracy of the landmarks to define the extent of various body regions (abdomen, pelvis etc) along the table length, which is relevant to typical medical scanning tasks. Figure 5 shows the error at various stages of the processing pipeline, averaged over the entire test dataset. Note that for depth sensors (based on structured light), the error in depth estimation increases quadratically with distance (red curve in the figure) and hence, the error in landmark estimation increases as well. However, even

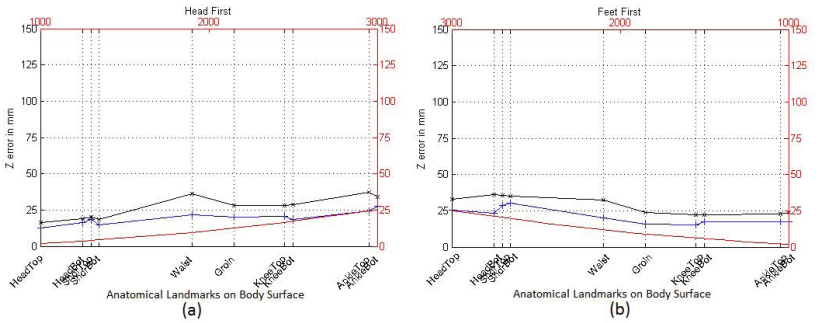

Fig. 5. Quantitative Results. (a) and (b) show the landmark estimation errors for head first and feet first poses respectively. The red curve indicates the error in mm in depth estimation from the depth sensor. Black and blue curves indicate the initial and optimized landmark detection results, respectively.

with the sensor noise, our algorithm still localizes the landmarks with less than 15 mm error for landmarks close to the gantry with error increasing only to 25 mm for the farthest landmarks.

4 Conclusion

A technical approach, which is able to obtain the 3D patient surface model from a single shot of a camera mounted on top of the scanner gantry is proposed. The technology is applicable to different types of medical scanner, such as CT, MR or PET. We validated the approach on real-world data and show robust and promising results. The proposed approach can potentially be used to provide several significant clinical benefits to existing scanning workflows, including time saving, radiation reduction, higher image quality, and real time guidance. Our future work includes capturing real patient camera data with their CT, MR, or PET data to further validate the performance for specific use cases as well as quantitatively evaluate the performance of the dense shape estimation, and to speed up the computation time to potentially update the model in real time.

References

1. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: SCAPE: shape completion and animation of people. ACM Trans. Graph. 24, 408–416 (2005)
2. Bauer, S., Wasza, J., Haase, S., Marosi, N., Hornegger, J.: Multi-modal surface registration for markerless initial patient setup in radiation therapy using microsoft's kinect sensor. In: IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 1175–1181 (November 2011)
3. Chen, Y., Medioni, G.: Object modeling by registration of multiple range images. In: IEEE International Conference on Robotics and Automation, vol. 3, pp. 2724–2729 (1991)
4. Frenzel, T.: Patient setup using a 3d laser surface scanning system. In: Dossel, O., Schlegel, W.C. (eds.) World Congress on Medical Physics and Biomedical Engineering. IFMBE Proceedings, vol. 25/1, pp. 217–220. Springer, Heidelberg (2009)
5. Rusu, R.B., Cousins, S.: 3D is here: Point cloud library (pcl). In: International Conference on Robotics and Automation. Shanghai, China (2011)
6. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1297–1304 (2011)
7. Sigal, L., Isard, M., Haussecker, H., Black, M.: Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. International Journal of Computer Vision 98(1), 15–48 (2012)
8. Tu, Z.: Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering. In: International Conference on Computer Vision (ICCV), vol. 2, pp. 1589–1596. IEEE (2005)