Towards Image-Based Modeling for Ambient Sensing

Julien Pansiot, Danail Stoyanov, Benny P.L. Lo & Guang-Zhong Yang
Royal Society/Wolfson Medical Image Computing Laboratory
Imperial College London, United Kingdom
e-mail: {jpansiot|dvs|benlo|gzy}@doc.ic.ac.uk

Abstract

The practical deployment of pervasive health monitoring requires a close integration of body sensor networks (BSNs) with intelligent ambient sensing. To this end, vision sensors with low cost and minimal power consumption provide an attractive means of activity tracking and capturing early signs of disease progression through changes in gait and posture. The purpose of this paper is to present an image-based modeling technique for generating a subject-specific simulation environment that allows systematic development of novel vision algorithms that can be implemented by low power video sensors with distributed processing and known ground-truth.

Keywords: *home care monitoring, camera network, simulation, shape from silhouette*

1. Introduction

In pervasive healthcare, extensive research is being directed towards the use of sensor networks for early disease detection, improved treatment compliance, and support for informal care provision. For BSNs, it becomes apparent that its full potential can only be realized by its effective integration with ambient sensing environment. While vision-based sensing technologies provide an intuitive, non-invasive and low cost solution for home monitoring, some major challenges need to be overcome as exemplified in gait and posture analysis research conducted by the computer vision community [10]. To address patient privacy issues, the concept of using non-appearance based techniques has been proposed [9] and existing research has also focused on developing subject-specific models and simulation environments to allow for systematic development of vision-based techniques [8]. The simulator is designed for distributed ambient intelligent sensing and can simulate different home layouts, camera views and subject activities. This permits an integrated design and implementation of algorithms related to activity tracking and distributed processing combined with rigorous testing with known ground truth data.

One of the key prerequisites of the simulation framework is the accurate representation of subject-specific behaviour. Our current approach for this purpose includes generating photo-realistic views of the virtual subjects. To achieve the high resolution required for photo-realism, video sequences of actors simulating different gaits and postures have been acquired from multiple view points to enable the use of image-based rendering techniques to generate novel views of the subject that are not covered by the physical vision sensors.

To obtain an arbitrary view of the subject, it is necessary to have a combination of knowledge about the 3D geometry of the cameras (both real and virtual) and the subject. This information may be obtained by using camera calibration [12] or from the actual video data by using self-calibration through feature matching. 3D information of the subject can also be determined directly from the video sequence by using image-based 3D reconstruction strategies based on structure-from-motion (SFM) [1] or space carving [3]. To achieve high fidelity rendering, the recovered 3D structure needs to be dense in order to generate views from virtual cameras from any distance. In this study, we have developed a Shape-From-Silhouette (SFS) method (also referred to as Image-Based Visual Hulls [6]) for 3D subject reconstruction that is resilient to camera resolution and relative camera placements. With SFS, the silhouettes of the subject are extracted in the video data and then back projected according to the camera parameters. To render a novel view from an arbitrary viewing angle and position, sampling of the ray intersection is performed and then shaded according to the initial colours obtained from the video data.

In the following sections, we present the details of the proposed framework for generating novel views of the subject for high fidelity virtual simulation. We will also show initial results obtained for subjects in a real camera network.

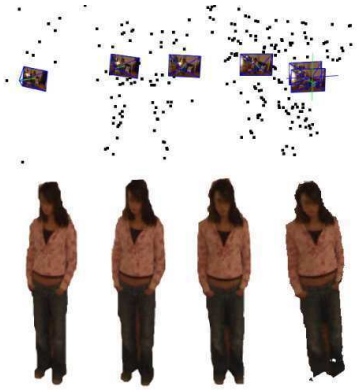

Figure 1. Extrinsic parameters estimated by the proposed method and a series of segmented silhouettes. The final silhouette is generated using image-based rendering.

2. 3D model sampling

In this study, the acquisition of the 3D model of the human body consists of background segmentation and 3D reconstruction. Background segmentation is performed by using a statistical background subtraction technique followed by morphological erosion/dilation filters to remove small isolated image features. Shadow removal is achieved by using the technique developed by Lo *et al* [4].

Calibration of the sensor nodes is performed by a two stage process consisting of determining the internal optics of each device and then finding the spatial configuration of the network. The intrinsic parameters of each camera are computed by using the calibration algorithm reported by Zhang [12]. With this approach, a calibration object is necessary but this is not inhibitive to a practical ubiquitous multi-camera network, where it is feasible to determine the intrinsic optics of each camera node before system installation. The extrinsic parameters, however, are site specific and need to be calculated individually for every network instance. For the ease of home deployment, we have developed a method for automatically determining the extrinsic parameters of the network based on the multi-view geometry between the camera nodes. A combination of gradient-based and SIFT keypoints [5] is used for establishing matching view relations. To solve for the rotation and translation parameters between nodes, we robustly calculate the essential matrix and decompose it to yield the valid configuration [7]. The dual feature strategy of the method favors nodes in various respective positions and the real-time capabilities of the method enable active feedback during network installation. Our initial experience without globally optimizing the resulting extrinsic network parameters suggests that the method can be used to avoid the use of calibration targets visible in multiple network nodes for system initialization. An example result of the estimated camera parameters is shown in Figure 1.

Once the 2D shapes of the object for the cameras are extracted, they are perspectively projected to their respective cameras, generating a generalised pyramid. By assuming a pinhole camera model, the intersection is then sampled by using a pyramidal set of rays whose apex is the desired point of view. The result at this stage of the process is highly dependent on the accuracy of both the intrinsic and extrinsic camera parameters. It is also important to note that the remaining set of rays is dependent on the sampling of the intersection.

Figure 2. An example of visual hull showing the original raw images and the modeled 3D data from a novel view angle.

3. Novel view rendering

In order to generate a photo-realistic view from the depth map, visibility detection is required to determine the line of sight and the final image shading of the object point. This is not trivial considering the fact that the object is not constrained in shape or form. In this study, we relied on the reverse projection of the sampled rays to the camera image plane. The distance from the camera centre to the surface of the rays gives the camera depth map. Note that the sampling rate of this depth map must be chosen according to that of the novel views, if the accuracy of the model is inadequate, the algorithm can detect extended visibility. This is particularly obvious at the edges of the model, where the rays can pass close to the surface.

For surface shading, there are several possible approaches that are applicable to the visual hull. Matusik [6] relied on a view-dependent technique which simply selects the reference camera whose angle to the desired target is the smallest. Another method is simply to choose the closest camera. Both of these are called *nearest neighbour interpolation* [11]. We also implemented a method that selects

the camera which is the most perpendicular to the surface as this usually means a better resolution. However, the quality of the result is highly dependent on the model and the camera configuration, and therefore should be used only in well-controlled environments. An example of a shaded image is shown in Figure 2.

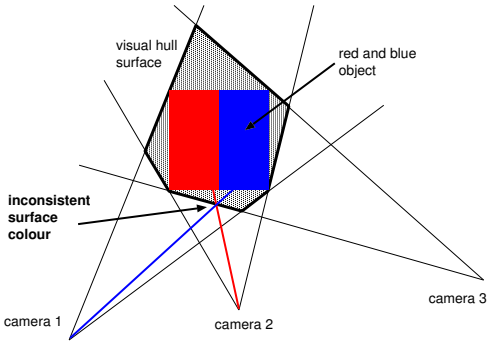

Figure 3. Visual Hull: how slightly inaccurate 3D model leads to shading errors.

Since it is not possible to obtain a perfectly sampled 3D model, the reprojection of the reference images can lead to gaps or overlaps as shown in Figure 3. Unlike techniques such as space carving [3], the pixels on the surface of a visual hull are not necessarily photo-consistent. For this reason, seams are likely to appear. In the final rendered images, this issue is tackled by Watson *et al* [11], but their method requires relatively large overlaps between reference images and could not be used in our context. A *stitching* algorithm was implemented to resolve this issue.

While the reference images are back-projected onto the model, consistency tracking is kept for the final coverage of every image. It is assumed that gaps and ovelaps are more likely to appear at the edges between two reference image reprojections. The *stitching* algorithm follows such edges and redefines the colours around them. For each point of the edge, a perpendicular vector crossing the edge is defined. The colours at each end of this vector are sampled and an linear gradient between these reference colours is written along the vector. An example is given in Figure 4.

One of the main applications of the above image-based rendering scheme is to generate canonical views for gait and posture classification. These synthetic images can be taken from a point of view that would remain static in the subject coordinate system, and thus greatly increase the sensitivity of algorithms such as the PCA and ICA for extracting the principal modes of the shape variation. To achieve this, we used an approximate model generated as above and combined with an oriented bounding box (OBB) [2] based on the eigenvectors of the covariance matrix of the sampling rays. The orientation of such a bounding box matches the

Figure 4. Stitching: initial image on top & stitched one below.

best shape of the cloud of points. Assuming that in most cases this bounding box will follow the general "natural" orientation of the body, we can rely on it to generate automatically six virtual points of view (front, back, top, bottom, left, right). Based on the above steps, six novel canonical views can be generated. A schematic diagram of the whole process is shown in Figure 5.

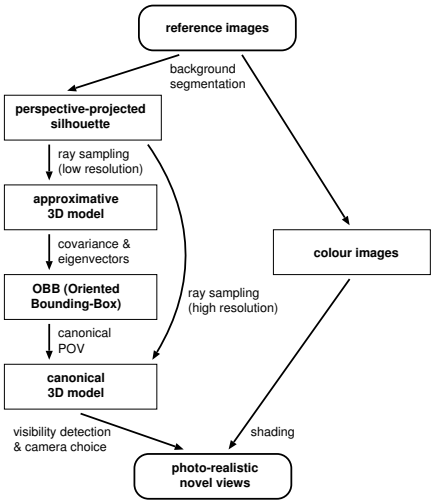

Figure 5. Generation of a canonical novel view.

It should be noted that some camera configuration can limit the validity of these views. Because the sampling of the visual hull can be highly inaccurate, visible artifacts (such as a large hump in the back of the subject) may lead to a mis-orientated bounding box.

4. Quantitative results

We use both photo-realism (self-consistency) and fidelity (consistency to the ground-truth data) measures for quantitative assessment of the results. To compare a novel image and a reference image, a scene was acquired with four

Figure 6. Reference vs. novel images.

Average silhouettes distance	4.57 pix. (1.21%)
Average silhouettes overlap	92%
Pearson correlation coeff.	0.85

Table 1. Comparison estimators.

cameras, but only three were used in the reconstruction process. The remaining one was kept as a reference image. The novel image was then computed from the view point of the reference camera and the difference between the synthesized and real images was compared. Two real and synthesized images are given in Figure 6. The images were cropped to the bounding box of the silhouettes to avoid an artificially high global correlation due to a large proportion of the background.

To provide a global measure of the error distribution, we computed the per-pixel Euclidean distance in the colour space between the two images and the result is shown in Figure 7 (a). It can be seen that the error is mostly concentrated on the sides of the silhouette, particularly on the upper body. The global shape, however, is only slightly affected. We also computed a map of the distance in hue shown in Figure 7 (b) to illustrate the accuracy of the colour rendering.

Finally, a number of quantitative measures were derived to evaluate the similarity between the reference image and the novel computed one as shown in Table 1. Given the importance of the silhouette in posture and gait analysis, we first computed the average distance between the two silhouettes. As the RMS is relatively sensitive to noise (in particular those generated by the reference camera), the *Pearson product-moment correlation coefficient* was prefered. It estimates the global correlation between the images.

5. Conclusion and future work

In this paper, we have demonstrated an image-based technique for generating novel views of the object from arbitrary viewing angles of the video sensors to augment the tracking capabilities of the BSNs. The qualitative and quantitative

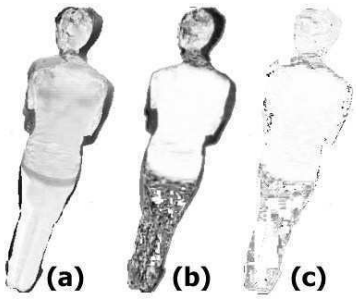

Figure 7. Comparison of the real image against a novel view: white means no difference, black means the maximum distance. (a) simple distance, (b) Pearson correlation coefficient, (c) hue.

comparisons between real reference images and synthesized images illustrate the potential value of the technique in generating photo-realistic views of the subject from any view angles. By the use of OBBs, canonical views can be derived, which can facilitate subject specific gait and posture analysis.

References

[1] C. Bregler and J. Malik. Tracking people with twists and exponential maps. *CVPR98*, 1998.

[2] S. Gottschalk, M. C. Lin, and D. Manocha. OBBTree: A hierarchical structure for rapid interference detection. *Computer Graphics*, 30(Annual Conference Series):171–180, 1996.

[3] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. Technical Report TR692, 1998.

[4] B. P. L. Lo and G.-Z. Yang. Neuro–fuzzy shadow filter. *European Conference on Computer Vision (ECCV) 2002 Part III*, pages 381–392, May 2002.

[5] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[6] W. Matusik. Image-based visual hulls. In *Master of Science Thesis*, 2001.

[7] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, June 2004.

[8] J. Pansiot, B. Lo, and G.-Z. Yang. A simulator for distributed ambient intelligence sensing. In *International Workshop on Wearable and Implantable Body Sensor Networks*, The IEE, page 119, London, April 2005.

[9] UbiSense. http://www.ubicare.org/projects-ubisense.shtml.

[10] J. Wang and S. Singh. Video analysis of human dynamics: a survey. *Real Time Imaging*, 2003.

[11] G. Watson, P. O'Brien, and M. Wright. Towards a perceptual method of blending for image-based models. *SIGRAD*, 2002.

[12] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2000.