Model-based Analysis of Multi-Video Data

Marcus Magnor and Christian Theobalt
Max-Planck-Institut für Informatik
Saarbrücken, Germany
magnor,theobalt@mpi-sb.mpg.de

Abstract

We describe a method to analyze multiple synchronized video streams by making use of a parameterized geometry model of the recorded object. By formulating the task of fitting the model to the images in terms of optimizing silhouette match, we are able to automatically and robustly capture the time-varying 3D pose of the object. To evaluate the energy functional, we exploit the fast image synthesis capabilities of a conventional PC graphics card. The use of an a-priori object model enables us to enforce kinematic constraints as well as temporal coherence, and we obtain a high-quality surface description as output. Suitably modified, the presented technique is also applicable to medical and other image analysis tasks if a parameterized, generic geometry model of the object of interest is available.

1. Introduction

When processing data collected by any type of imaging technique, often substantial knowledge about what has been recorded is available a-priori. Modeled in suitable form, this information may be exploitable to efficiently constrain image analysis and interpretation processing.

This paper presents a method to robustly capture the motion of a dynamic object from multiple synchronized video recordings, given an adaptable geometry model of the object. The use of a parameterized model, consisting of multiple rigid body segments, enables us to enforce kinematic constraints as defined by the nature of the object, and to impose temporal coherence. Besides ensuring a physically plausible evolution of the object's pose, these constraints also efficiently restrict parameter search space, leading to a robust, automatic optical motion capture algorithm.

In the following section, we briefly reflect on related work regarding model-based analysis of image data. The implementation of our analysis-by-synthesis approach exploiting PC graphics hardware is outlined in Sect. 3. We go on to describe in Sect. 4 how a geometry model of a human is automatically matched to an actor recorded with a handful of synchronized video cameras. Sect. 5 outlines how our silhouette-based algorithm can be parallelized in order to attain faster performance. Results of our optical motion capture sytem are presented in Sect. 6 before we conclude in Sect. 7 with an outlook on potential other applications for the described technique.

2. Related Work

Methods to exploit a-priori 3D geometry information for image analysis purposes have been investigated by various researchers [16]. Here, we concentrate on algorithms that make use of a parameterized geometry model to analyze temporally varying scene content, recorded by synchronized video cameras.

Linearized reconstruction from optical flow has been employed to determine facial animation parameters [6] as well as body pose [8, 2] for suitable geometry models. If multiple video cameras are available, reconstructing 3D geometry first and then fitting a coarse human body representation to the geometry allows qualitative capture of human motion [1, 5, 13]. The human body model can be matched to video image content also if dense depth maps and object silhouettes are available [15]. These approaches are tailored specifically to their application, require additional, error-prone pre-processing steps, or deliver only approximate results.

In contrast, we demonstrate how a high-quality, generic geometry model can be fit robustly to video image data using only object silhouettes [4]. Matching 2D image contours to projected object geometry outline has been proven useful for estimating camera parameters [3, 11, 7, 14, 12], and silhouette area has been used to register still photographs to scanned 3D object geometry [9, 10].

In the following, we describe in detail a silhouette-based fitting method suitable for adapting any arbitrary geometry model to multi-video data while simultaneously enforcing the object's kinematic constraints and guaranteeing a temporally coherent evolution.

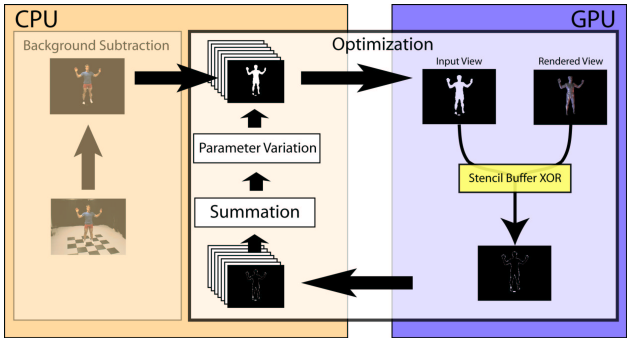

Figure 1. Model-to-silhouette adaptation: The binary, segmented video images are XORed with the rendered model. The number of remaining pixels is minimized by optimizing model parameter values.

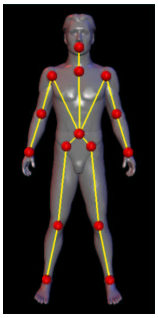

Figure 2. The geometry model consists of 16 rigid body parts. Each body part can be independently scaled and its surface deformed, while 35 joint parameters determine actual body pose.

3. Silhouette-based Model Fitting

To compare the model of an object's 3D geometry to its appearance in multiple video images, we make use of the object's image silhouettes. Segmentation of the silhouette in the video images can be achieved in various ways, e.g. by pre-recording the static background and appropriate thresholding each pixel. We render the model from all camera perspectives and compare the model silhouette area to the segmented video images by performing an exclusive-or operation between each image's binary silhouette mask and the corresponding rendered model outline and counting the number of remaining pixels, Fig. 1. The task of finding the best parameter values matching the model to the video images then becomes an optimization problem of minimizing the sum of set pixels.

Using image silhouettes to compare model pose to actual object appearance has numerous advantages:

- Silhouettes can be very easily and robustly extracted,
- they provide a large number of pixels to overdetermine the parameter search,
- silhouettes of the geometry model can be rendered very efficiently on modern graphics hardware, and
- the XOR energy functional can be implemented entirely on graphics hardware. Up to 8 binary foreground masks from the segmented video images and the corresponding rendered model can be XORed simultaneously using the graphics card's stencil buffer. The number of set pixels is determined by calculating the (binary) histogram.

On the CPU, a standard optimization algorithm (e.g. Powell's method) iteratively alters the pose parameter values to minimize the energy functional, as evaluated on the graphics card.

To avoid local minima and to obtain accurate model parameter values, we compose our model hierarchically and vary only subsets of model parameters simultaneously. In addition, we perform a grid search to initialize our optimization routine, i.e., a small number of candidate values around predicted parameter values are considered and the energy function evaluated. This accelerates convergence and has been found to efficiently avoid local minima.

4 An Example: Optical Motion Capture

To illustrate how our silhouette-based algorithm works in practice, we consider the task of capturing the complex motion of a human jazz dance performance from multiple synchronized video recordings. A publically available VRML geometry model of a human body serves as our generic geometry model, Fig. 2. The model consists of 16 rigid body segments, one for the upper and lower torso, neck, and head, and pairs for the upper arms, lower arms, hands, upper legs, lower legs and feet. In total, more than 21000 triangles make up our human body surface. All body segments are connected by a kinematic chain, resembling the anatomy of the human skeleton. 17 joints with a total of 35 joint parameters define the pose of our virtual character. Since our model parameterization is based on human anatomy, we can incorporate additional knowledge about what a human dancer can and cannot do, thus constraining the outcome of our optimization to only plausible results.

Since we start out with a generic body model, the initial geometry will not have the same proportions as its human counterpart. Therefore, the model dimensions must be adaptable to match the size and proportions of the human dancer. To this purpose, each rigid body segment can

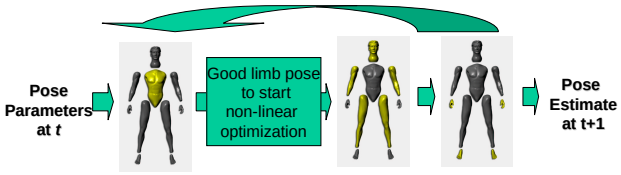

Figure 3. The body parts are fit to the images in hierarchical order: The torso first, then the upper arms, legs and the head, and finally the lower arms, legs, hands and feet. In conjunction with a small exhaustive grid search around some parameters' predicted value for optimization intialization, local energy functional minima are efficiently avoided.

be scaled along its underlying anatomical bone structure. In addition, a local Bézier parameterization of the triangle mesh allows us to non-uniformly deform each body segment's surface by tweaking 16 control parameters per segment. This way, we can closely match the stature of the body model to the built of the human individual in the video images.

For model initialization, the dancer stands still for a short moment to have his silhouette recorded from all camera perspectives. To adapt model proportions to the human character, first, only the body torso is considered and its position and orientation is approximately found by maximizing the overlap with all silhouettes. Then the pose of arms, legs and head are recovered by first rendering each limb in a number of orientations and selecting the best match as intialization for a refined optimization. After the generic model has been approximately fit, the uniform scaling parameters of each body segment are adjusted. The algorithm then alternates between optimizing joint parameters and body segment scaling parameters until it has converged to the true pose and proportions of the person. Once the correct body pose and body segment proportions have been found, the Bézier control parameters of all body segments are optimized to match each segment's outline to the recorded silhouettes.

In the following, only the 35 joint parameters are varied to capture the pose of the dancer over time. With our individualized geometry model, we determine the joint parameter values for each time instant such that our model closely follows the motion of the human character. Model parameter estimation is performed in hierarchical order with respect to their impact on silhouette appearance and their position in the model's kinematic chain, Fig 3. To find the correct joint parameter values at a time instant $t + 1$ from the current pose, we make use of the parameter values' history and predict their new values from their recent rate of

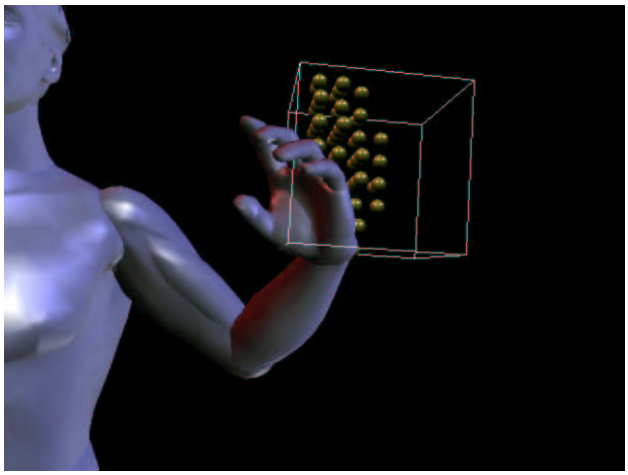

Figure 4. Grid search initialization: To avoid local minima, optimization of some joint parameters is preceded by testing a few parameter values in the vicinity of the predicted position.

change. The model is hierarchically updated: First, position and orientation of the torso are varied to find the 3D location of the body. Next, upper arms, thighs and head are considered. Finally, the lower extremities as well as hands and feet are regarded.

To make sure we do not end up in a local minimum, we additionally perform a grid search in parameter space for some joints, Fig. 4: A small number of candidate parameter values around the predicted new value are considered and the energy function evaluated. The optimization routine is then intialized with the best set of parameter values of the grid search.

One last constraint we have to enforce is the avoidance of inter-penetrations of different body segments. By testing against all segments' bounding boxes during our grid search, we ensure that all parameter values correspond to a plausible, non-interpenetrating body model.

5. Parallelized Silhouette Fitting

Our silhouette-matching approach is an iterative algorithm that converges robustly towards the best set of joint parameter values. Employing Powell's optimization method on a 1.8 GHz PC and evaluating the energy functional on a GeForce3 graphics card, human pose analysis takes between 8 and 14 seconds per time instant. However, our algorithm is easily parallelizable such that multiple computers and graphics cards can work simultaneously on estimating pose parameters [17]. Since the root node of the kinematic chain, the lower torso, branches into 5 sep-

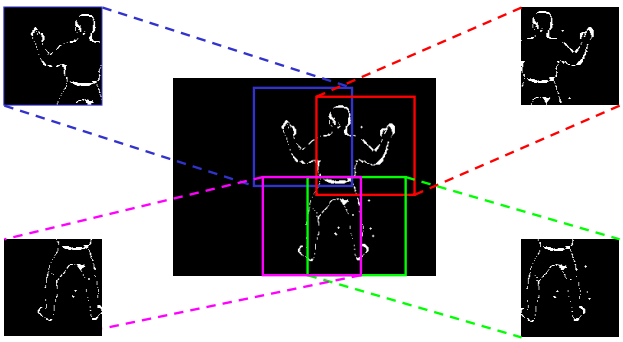

Figure 5. Parallelized implementation: Each body extremity is evaluated on one PC / graphics board.

arate body parts, the poses of all four extremities and the head can be determined independently once torso position and orientation are known, Fig. 5. Five personal computers with graphics cards can therefore work on estimating the correct model parameters in parallel. After the server PC has determined the 6 degrees of freedom of torso position and orientation, each of the four client PCs and the server each optimize only one arm, leg, and the head. The video images are windowed around each body part's approximate position and are transferred via conventional ethernet to the corresponding client. To reduce the influence of other body parts during separate body limb optimization, the entire geometry model is rendered once without the body segment in question, and all covered silhouette pixels are discarded, Fig. 6. While this masking operation cannot eliminate all pixels stemming from other body parts, it nevertheless efficiently reduces the number of non-optimized pixels and raises the score of those pixels belonging to the body part whose pose is being optimized. This way, only that body segment is locally rendered whose pose parameters are to be found during optimization, and it is compared only to the non-masked image silhouette pixels. This also reduces the number of triangles to be rendered.

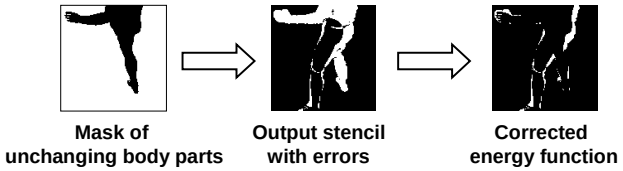

Figure 6. Body parts not optimized by one client computer are masked prior to optimization to exclude them during energy functional evaluation.

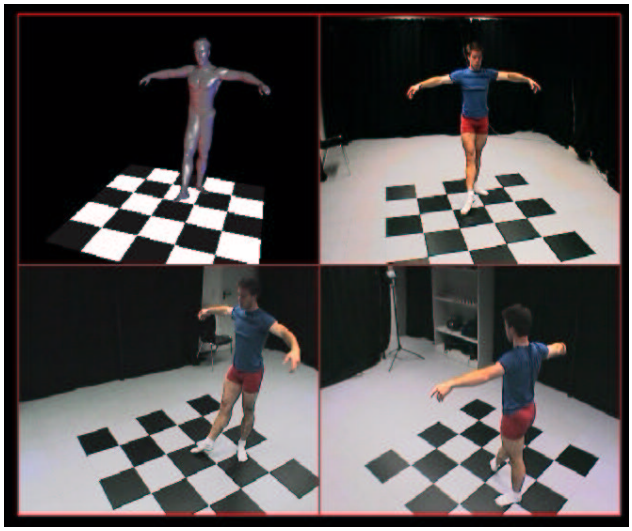

Figure 7. Fitted geometry model and three corresponding input video images of its human counterpart.

The parallelized implementation of our silhouette-based model fitting algorithm currently reduces the time it takes to estimate full body pose by a factor of 10 down to less than 2 seconds per frame.

6. Results

Given eight input video streams, our model-based analysis approach robustly captures the motion of the dancer [1], Figs. 7, 8. On a single PC with an nVidia^{TM}GeForce3 graphics card, pose estimation takes between 8 and 14 seconds per time instant. When optimizing each body limb on a separate PC, pose estimation time reduces to less than 2 seconds per time step.

While we have presented here a non-invasive human motion capture system, it should be noted that the model-based video analysis approach is generally applicable to any arbitrary object that can be represented by a number of interconnected rigid body segments, or by a suitably parameterized non-rigid object model.

Exploiting the graphics card's stencil buffer, the error functional can be evaluated on any modern graphics card. With an nVidia GeForce3 card, more than 100 evaluations can be performed per second taking into account 8 complete camera images simultaneously. Only the Powell optimization algorithm needs to run on the CPU.

One valuable advantage of our model-based object analysis approach is the low-dimensional parameter search

[1] movie samples can be found at *www.grovis.de/fvv*

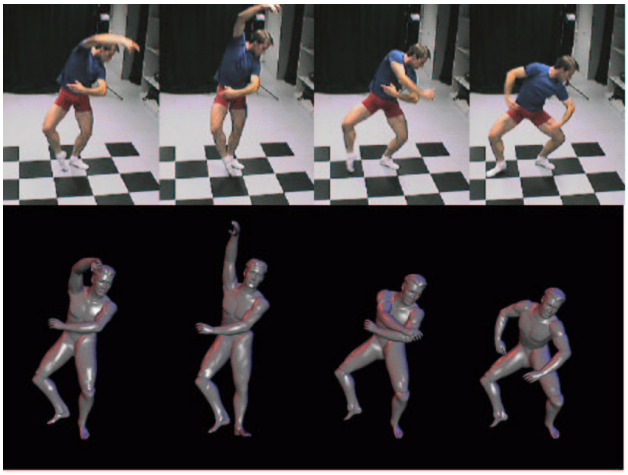

Figure 8. Model-based multi-video analysis of a jazz dance performance.

space, since out model parameterization provides only a few dozen degrees of freedom to match model and image silhouettes. In addition, constraints on parameter values are easily enforced by making sure that during optimization, all parameter values stay within their anatomically plausible range. Finally, temporal coherence is maintained by allowing only a maximal change in magnitude for each parameter from one time step to the next.

7. Conclusions

We have presented a model-based video analysis method that relies solely on object silhouette information. In an analysis-by-synthesis loop, we exploit the fast rendering capabilities of conventional PC graphics hardware to match model outline to image silhouettes. The approach robustly determines 3D object pose, taking into account kinematic constraints and maintaining temporal coherence.

The algorithm is potentially applicable to a wide range of image analysis and interpretation tasks. Wherever a parameterized geometry model of a recorded object is available and object silhouette can be (even coarsely) determined, silhouette-based analysis-by-synthesis can be applied. In addition, the approach can be easily extended to 3D volume data. For example, in medical image analysis the surface of a parameterized organ model can be matched to three-dimensional CT or MRI data to identify and further process the organ's voxels, e.g., for later automated diagnosis.

References

[1] A. Bottino and A. Laurentini. A silhouette based technique for the reconstruction of human movement. *Journal of Computer Vision and Image Understanding*, 83:79–95, 2001.

[2] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proc. of CVPR 98*, pages 8–15, 1998.

[3] L. Brunie, S. Lavallée, and R. Szeliski. Using Force Fields Derived from 3D Distance Maps for Inferring the Attitude of a 3D Rigid Object. In *Proceedings of Computer Vision (ECCV '92)*, volume 588, pages 670–675, Mai 1992.

[4] J. Carranza, C. Theobalt, M. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. *Proc. ACM Siggraph'03*, San Diego, USA, July 2003.

[5] K. Cheung, T. Kanade, J.-Y. Bouguet, and M. Holler. A real time system for robust 3D voxel reconstruction of human motions. In *Proc. of CVPR*, volume 2, pages 714 – 720, June 2000.

[6] P. Eisert, W. T., and B. Girod. Model-aided coding : A new approach to incorporate facial animation into motion-compensated video coding. *IEEE Trans. Circuits and Systems for Video Technology*, 10(3):244–258, 2000.

[7] H. H. S. Ip and L. Yin. Constructing a 3D individualized head model from two orthogonal views. *The Visual Computer*, 12(5):254–268, 1996.

[8] R. Koch. Dynamic 3D scene analysis through synthesis feedback control. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(6):556–568, 1993.

[9] H. Lensch, W. Heidrich, and H.-P. Seidel. Automated texture registration and stitching for real world models. In B. A. Barsky, Y. Shinagawa, and W. Wang, editors, *Proceedings of the 8th Pacific Conference on Computer Graphics and Applications (PG-00)*, pages 317–326, Hong Kong, China, October 2000. IEEE Computer Society.

[10] H. P. A. Lensch, W. Heidrich, and H.-P. Seidel. A silhouette-based algorithm for texture registration and stitching. *Graphical Models*, 63(4):245–262, July 2001.

[11] D. G. Lowe. Fitting Parameterized Three-Dimensional Models to Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(5):441–450, 1991.

[12] K. Matsushita and T. Kaneko. Efficient and Handy Texture Mapping on 3D Surfaces. *Computer Graphics Forum*, 18(3):349–358, September 1999.

[13] I. Mikić, M. Triverdi, E. Hunter, and P. Cosman. Articulated body posture estimation from multicamera voxel data. In *Proc. of CVPR*, 2001.

[14] P. J. Neugebauer and K. Klein. Texturing 3D Models of Real World Objects from Multiple Unregistered Photographic Views. *Computer Graphics Forum*, 18(3):245–256, September 1999.

[15] R. Plaenkers and P. Fua. Tracking and modeling people in video sequences. *Journal of Computer Vision and Image Understanding*, 81(3):285–302, March 2001.

[16] A. Pope. Model-based object recognition - A survey of recent research. Technical Report TR-94-04, Dept. Computer Science, Univ. British Columbia, Jan. 1994.

[17] C. Theobalt, J. Carranza, M. Magnor, and H.-P. Seidel. A parallel framework for silhouette-based human motion capture. *Proc. Vision, Modeling, and Visualization (VMV-2003)*, Munich, Germany, Nov. 2003.