Ahmed Elgammal
Bodo Rosenhahn
Reinhard Klette (Eds.)

LNCS 4814

Human Motion – Understanding, Modeling, Capture and Animation

Second Workshop, Human Motion 2007
Rio de Janeiro, Brazil, October 2007
Proceedings

Springer

Ahmed Elgammal Bodo Rosenhahn
Reinhard Klette (Eds.)

Human Motion – Understanding, Modeling, Capture and Animation

Second Workshop, Human Motion 2007
Rio de Janeiro, Brazil, October 20, 2007
Proceedings

Springer

Volume Editors

Ahmed Elgammal
Rutgers State University
Department of Computer Science
110 Frelinghuysen Road, Piscataway, NJ 08854-8019, USA
E-mail: elgammal@cs.rutgers.edu

Bodo Rosenhahn
Max Planck Center for Visual Computing and Communication
Stuhlsatzhausenweg 85, 66123 Saarbrücken, Germany
E-mail: rosenhahn@mpi-sb.mpg.de

Reinhard Klette
The University of Auckland
Computer Science Department
Private Bag 92019, Auckland Mail Center, Auckland 1142, New Zealand
E-mail: r.klette@auckland.ac.nz

Library of Congress Control Number: Applied for

CR Subject Classification (1998): I.2.9, I.2.10, I.4.8, I.7.5, I.3.7

LNCS Sublibrary: SL 6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

ISSN	0302-9743
ISBN-10	3-540-75702-3 Springer Berlin Heidelberg New York
ISBN-13	978-3-540-75702-3 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12175420 06/3180 5 4 3 2 1 0

Lecture Notes in Computer Science 4814

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Lecture Notes in Computer Science

Sublibrary 6: Image Processing, Computer Vision, Pattern Recognition, and Graphics

Vol. 4814: A. Elgammal, B. Rosenhahn, R. Klette (Eds.), Human Motion – Understanding, Modeling, Capture and Animation. X, 329 pages. 2007.

Vol. 4778: S.K. Zhou, W. Zhao, X. Tang, S. Gong (Eds.), Analysis and Modeling of Faces and Gestures. X, 305 pages. 2007.

Vol. 4738: A. Paiva, R. Prada, R.W. Picard (Eds.), Affective Computing and Intelligent Interaction. XVIII, 781 pages. 2007.

Vol. 4729: F. Mele, G. Ramella, S. Santillo, F. Ventriglia (Eds.), Advances in Brain, Vision, and Artificial Intelligence. XVI, 618 pages. 2007.

Vol. 4713: F.A. Hamprecht, C. Schnörr, B. Jähne (Eds.), Pattern Recognition. XIII, 560 pages. 2007.

Vol. 4679: A.L. Yuille, S.-C. Zhu, D. Cremers, Y. Wang (Eds.), Energy Minimization Methods in Computer Vision and Pattern Recognition. XII, 494 pages. 2007.

Vol. 4678: J. Blanc-Talon, W. Philips, D. Popescu, P. Scheunders (Eds.), Advanced Concepts for Intelligent Vision Systems. XXIII, 1100 pages. 2007.

Vol. 4673: W.G. Kropatsch, M. Kampel, A. Hanbury (Eds.), Computer Analysis of Images and Patterns. XX, 1006 pages. 2007.

Vol. 4642: S.-W. Lee, S.Z. Li (Eds.), Advances in Biometrics. XX, 1216 pages. 2007.

Vol. 4633: M. Kamel, A. Campilho (Eds.), Image Analysis and Recognition. XII, 1312 pages. 2007.

Vol. 4584: N. Karssemeijer, B. Lelieveldt (Eds.), Information Processing in Medical Imaging. XX, 777 pages. 2007.

Vol. 4569: A. Butz, B. Fisher, A. Krüger, P. Olivier, S. Owada (Eds.), Smart Graphics. IX, 237 pages. 2007.

Vol. 4538: F. Escolano, M. Vento (Eds.), Graph-Based Representations in Pattern Recognition. XII, 416 pages. 2007.

Vol. 4522: B.K. Ersbøll, K.S. Pedersen (Eds.), Image Analysis. XVIII, 989 pages. 2007.

Vol. 4485: F. Sgallari, A. Murli, N. Paragios (Eds.), Scale Space and Variational Methods in Computer Vision. XV, 931 pages. 2007.

Vol. 4478: J. Martí, J.M. Benedí, A.M. Mendonça, J. Serrat (Eds.), Pattern Recognition and Image Analysis, Part II. XXVII, 657 pages. 2007.

Vol. 4477: J. Martí, J.M. Benedí, A.M. Mendonça, J. Serrat (Eds.), Pattern Recognition and Image Analysis, Part I. XXVII, 625 pages. 2007.

Vol. 4472: M. Haindl, J. Kittler, F. Roli (Eds.), Multiple Classifier Systems. XI, 524 pages. 2007.

Vol. 4466: F.B. Sachse, G. Seemann (Eds.), Functional Imaging and Modeling of the Heart. XV, 486 pages. 2007.

Vol. 4418: A. Gagalowicz, W. Philips (Eds.), Computer Vision/Computer Graphics Collaboration Techniques. XV, 620 pages. 2007.

Vol. 4417: A. Kerren, A. Ebert, J. Meyer (Eds.), Human-Centered Visualization Environments. XIX, 403 pages. 2007.

Vol. 4391: Y. Stylianou, M. Faundez-Zanuy, A. Esposito (Eds.), Progress in Nonlinear Speech Processing. XII, 269 pages. 2007.

Vol. 4370: P.P. Lévy, B. Le Grand, F. Poulet, M. Soto, L. Darago, L. Toubiana, J.-F. Vibert (Eds.), Pixelization Paradigm. XV, 279 pages. 2007.

Vol. 4358: R. Vidal, A. Heyden, Y. Ma (Eds.), Dynamical Vision. IX, 329 pages. 2007.

Vol. 4338: P.K. Kalra, S. Peleg (Eds.), Computer Vision, Graphics and Image Processing. XV, 965 pages. 2006.

Vol. 4319: L.-W. Chang, W.-N. Lie (Eds.), Advances in Image and Video Technology. XXVI, 1347 pages. 2006.

Vol. 4292: G. Bebis, R. Boyle, B. Parvin, D. Koracin, P. Remagnino, A. Nefian, G. Meenakshisundaram, V. Pascucci, J. Zara, J. Molineros, H. Theisel, T. Malzbender (Eds.), Advances in Visual Computing, Part II. XXXII, 906 pages. 2006.

Vol. 4291: G. Bebis, R. Boyle, B. Parvin, D. Koracin, P. Remagnino, A. Nefian, G. Meenakshisundaram, V. Pascucci, J. Zara, J. Molineros, H. Theisel, T. Malzbender (Eds.), Advances in Visual Computing, Part I. XXXI, 916 pages. 2006.

Vol. 4245: A. Kuba, L.G. Nyúl, K. Palágyi (Eds.), Discrete Geometry for Computer Imagery. XIII, 688 pages. 2006.

Vol. 4241: R.R. Beichel, M. Sonka (Eds.), Computer Vision Approaches to Medical Image Analysis. XI, 262 pages. 2006.

Vol. 4225: J.F. Martínez-Trinidad, J.A. Carrasco Ochoa, J. Kittler (Eds.), Progress in Pattern Recognition, Image Analysis and Applications. XIX, 995 pages. 2006.

Vol. 4191: R. Larsen, M. Nielsen, J. Sporring (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006, Part II. XXXVIII, 981 pages. 2006.

Vol. 4190: R. Larsen, M. Nielsen, J. Sporring (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006, Part I. XXXVVIII, 949 pages. 2006.

Vol. 4179: J. Blanc-Talon, W. Philips, D. Popescu, P. Scheunders (Eds.), Advanced Concepts for Intelligent Vision Systems. XXIV, 1224 pages. 2006.

Vol. 4174: K. Franke, K.-R. Müller, B. Nickolay, R. Schäfer (Eds.), Pattern Recognition. XX, 773 pages. 2006.

Vol. 4170: J. Ponce, M. Hebert, C. Schmid, A. Zisserman (Eds.), Toward Category-Level Object Recognition. XI, 618 pages. 2006.

Vol. 4153: N. Zheng, X. Jiang, X. Lan (Eds.), Advances in Machine Vision, Image Processing, and Pattern Analysis. XIII, 506 pages. 2006.

Vol. 4142: A. Campilho, M. Kamel (Eds.), Image Analysis and Recognition, Part II. XXVII, 923 pages. 2006.

Vol. 4141: A. Campilho, M. Kamel (Eds.), Image Analysis and Recognition, Part I. XXVIII, 939 pages. 2006.

Vol. 4122: R. Stiefelhagen, J.S. Garofolo (Eds.), Multimodal Technologies for Perception of Humans. XII, 360 pages. 2007.

Vol. 4109: D.-Y. Yeung, J.T. Kwok, A. Fred, F. Roli, D. de Ridder (Eds.), Structural, Syntactic, and Statistical Pattern Recognition. XXI, 939 pages. 2006.

Vol. 4091: G.-Z. Yang, T. Jiang, D. Shen, L. Gu, J. Yang (Eds.), Medical Imaging and Augmented Reality. XIII, 399 pages. 2006.

Vol. 4073: A. Butz, B. Fisher, A. Krüger, P. Olivier (Eds.), Smart Graphics. XI, 263 pages. 2006.

Vol. 4069: F.J. Perales, R.B. Fisher (Eds.), Articulated Motion and Deformable Objects. XV, 526 pages. 2006.

Vol. 4057: J.P.W. Pluim, B. Likar, F.A. Gerritsen (Eds.), Biomedical Image Registration. XII, 324 pages. 2006.

Vol. 4046: S.M. Astley, M. Brady, C. Rose, R. Zwiggelaar (Eds.), Digital Mammography. XVI, 654 pages. 2006.

Vol. 4040: R. Reulke, U. Eckardt, B. Flach, U. Knauer, K. Polthier (Eds.), Combinatorial Image Analysis. XII, 482 pages. 2006.

Vol. 4035: T. Nishita, Q. Peng, H.-P. Seidel (Eds.), Advances in Computer Graphics. XX, 771 pages. 2006.

Vol. 3979: T.S. Huang, N. Sebe, M.S. Lew, V. Pavlović, M. Kölsch, A. Galata, B. Kisačanin (Eds.), Computer Vision in Human-Computer Interaction. XII, 121 pages. 2006.

Vol. 3954: A. Leonardis, H. Bischof, A. Pinz (Eds.), Computer Vision – ECCV 2006, Part IV. XVII, 613 pages. 2006.

Vol. 3953: A. Leonardis, H. Bischof, A. Pinz (Eds.), Computer Vision – ECCV 2006, Part III. XVII, 649 pages. 2006.

Vol. 3952: A. Leonardis, H. Bischof, A. Pinz (Eds.), Computer Vision – ECCV 2006, Part II. XVII, 661 pages. 2006.

Vol. 3951: A. Leonardis, H. Bischof, A. Pinz (Eds.), Computer Vision – ECCV 2006, Part I. XXXV, 639 pages. 2006.

Vol. 3948: H.I. Christensen, H.-H. Nagel (Eds.), Cognitive Vision Systems. VIII, 367 pages. 2006.

Vol. 3926: W. Liu, J. Lladós (Eds.), Graphics Recognition. XII, 428 pages. 2006.

Vol. 3872: H. Bunke, A.L. Spitz (Eds.), Document Analysis Systems VII. XIII, 630 pages. 2006.

Vol. 3852: P.J. Narayanan, S.K. Nayar, H.-Y. Shum (Eds.), Computer Vision – ACCV 2006, Part II. XXXI, 977 pages. 2006.

Vol. 3851: P.J. Narayanan, S.K. Nayar, H.-Y. Shum (Eds.), Computer Vision – ACCV 2006, Part I. XXXI, 973 pages. 2006.

Vol. 3832: D. Zhang, A.K. Jain (Eds.), Advances in Biometrics. XX, 796 pages. 2005.

Vol. 3736: S. Bres, R. Laurini (Eds.), Visual Information and Information Systems. XI, 291 pages. 2006.

Vol. 3667: W.J. MacLean (Ed.), Spatial Coherence for Visual Motion Analysis. IX, 141 pages. 2006.

Vol. 3417: B. Jähne, R. Mester, E. Barth, H. Scharr (Eds.), Complex Motion. X, 235 pages. 2007.

Vol. 2396: T.M. Caelli, A. Amin, R.P.W. Duin, M.S. Kamel, D. de Ridder (Eds.), Structural, Syntactic, and Statistical Pattern Recognition. XVI, 863 pages. 2002.

Vol. 1679: C. Taylor, A. Colchester (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI'99. XXI, 1240 pages. 1999.

Preface

This LNCS volume contains the papers presented at the second Workshop on Human Motion Understanding, Modeling, Capture and Animation, which took place on October 20th, 2007, accompanying the 11th IEEE International Conference on Computer Vision in Rio de Janeiro, Brazil.

In total, 38 papers were submitted to this workshop, of which 22 papers were accepted. We were careful to ensure a high standard of quality when selecting the papers. All submissions were double-blind reviewed by at least two experts. Out of the 22 accepted papers, 10 were selected for oral presentation and 12 for posters. We thank the authors of the accepted papers for taking the reviewers' comments into account in the final published versions of their papers. We thank all of the authors who submitted their work, and we trust that the reviewers' comments have been of value for their research activities.

The accepted papers reflect the state of the art in the field and cover various topics related to human motion tracking and analysis. The papers in this volume have been classified into three categories based on the topics they cover: human motion capture and pose estimation, body and limb tracking and segmentation, and activity recognition.

It was a special honor to have Prof. Demetri Terzopulos (University of California, Los Angeles) as the invited speaker at the workshop. We are especially grateful to the members of the Program Committee for their remarkable efforts and the quality of their timely reviews. The organization of this event would not have been possible without the effort and the enthusiasm of several people, and we thank all who contributed.

October 2007

Ahmed Elgammal
Bodo Rosenhahn
Reinhard Klette

Organization

Organizing Committee

Program Chairs
Ahmed Elgammal (Rutgers University, USA)
Bodo Rosenhahn (Max-Planck Institute for Computer Science, Germany)
Reinhard Klette (The University of Auckland, New Zealand)

Program Committee

Michael Black (Brown University, USA)
Richard Bowden (University of Surrey, UK)
Thomas Brox (University of Bonn, Germany)
Stefan Carlsson (KTH Royal Institute of Technology, Sweden)
Larry Davis (The University of Maryland, College Park, USA)
Leo Dorst (University of Amsterdam, The Netherlands)
Pascal Fua (Ecole Polytechnique Fédérale de Lausanne, Switzerland)
David Fleet (University of Toronto, Canada)
Vaclav Hlavac (Czech Technical University, Czech Republic)
Jessica K. Hodgins (Carnegie Mellon University, USA)
Atsushi Imiya (Chiba University, Japan)
Reinhard Koch (Christian-Albrechts-University of Kiel, Germany)
Volker Krüger (Aalborg University, Denmark)
Marcus Magnor (TU Braunschweig, Germany)
Dimitris Metaxas (Rutgers University, USA)
Meinard Müller (University of Bonn, Germany)
Lars Muendermann (Stanford University, USA)
Michael Neff (University of California, Davis, USA)
Ramakant Nevatia (University of Southern California, USA)
Vladimir Pavlovic (Rutgers University, USA)
Fatih Porikli (Mitsubishi Electric Research Laboratories, USA)
Stan Sclaroff (Boston University, USA)
Hans-Peter Seidel (Max-Planck Institute for Computer Science, Germany)
Cristian Sminchisescu (Bonn University, Germany)
Gerald Sommer (Christian-Albrechts-Universität zu Kiel)
Demetri Terzopoulos (University of California, Los Angeles, USA)
Matthias Teschner (University of Freiburg, Germany)
Christian Theobalt (Stanford University, USA)
Matthew Turk (University of California, Santa Barbara, USA)
Jian J. Zhang (Bournemouth University, UK)

Additional Reviewers	Nils Hasler (Max-Planck Institute for Computer Science, Germany) Chan Su Lee (Rutgers University, USA) Christian Schmaltz (Saarland University, Germany) Martin Sunkel (Max-Planck Institute for Computer Science, Germany)
Editorial Assistant	Chan Su Lee (Rutgers University, USA)

Table of Contents

Motion Capture and Pose Estimation

Body and Limb Tracking and Segmentation

Activity Recognition

Marker-Less 3D Feature Tracking for Mesh-Based Human Motion Capture

Edilson de Aguiar[1], Christian Theobalt[2], Carsten Stoll[1], and Hans-Peter Seidel[1]

[1] MPI Informatik, Germany
[2] Stanford University, USA
{edeaguia, stoll, hpseidel}@mpi-inf.mpg.de, theobalt@cs.stanford.edu

Abstract. We present a novel algorithm that robustly tracks 3D trajectories of features on a moving human who has been recorded with multiple video cameras. Our method does so without special markers in the scene and can be used to track subjects wearing everyday apparel. By using the paths of the 3D points as constraints in a fast mesh deformation approach, we can directly animate a static human body scan such that it performs the same motion as the captured subject. Our method can therefore be used to directly animate high quality geometry models from unaltered video data which opens the door to new applications in motion capture, 3D Video and computer animation. Since our method does not require a kinematic skeleton and only employs a handful of feature trajectories to generate lifelike animations with realistic surface deformations, it can also be used to track subjects wearing wide apparel, and even animals. We demonstrate the performance of our approach using several captured real-world sequences, and also validate its accuracy.

1 Introduction

Nowadays, generating realistic and lifelike animated characters from captured real-world motion sequences is still a hard and time-consuming task. Traditionally, marker-based optical motion capture systems [1] reconstruct the motion of a moving subject by measuring the 3D trajectories of optical beacons attached to her body. The optical markers are then mapped to a kinematic skeleton structure [2]. Marker-free methods also exist that are able to measure human motion in terms of a kinematic skeleton without any intrusion into the scene. Thereafter, the model geometry and the skeleton need to be connected such that the surface deforms realistically with the body motion by specifying the influence of each bone on both rigid and non-rigid surface deformation [3].

Stepping directly from a captured real-world sequence to the corresponding realistic moving character is still challenging. Several methods in the literature are able to partly solve this problem. Since marker-based and marker-free motion capture systems measure the motion in terms of a kinematic skeleton, they have to be combined with other scanning technologies to capture the time-varying shape of the human body surface [4,5,6]. However, dealing with people

A. Elgammal et al. (Eds.): Human Motion 2007, LNCS 4814, pp. 1–15, 2007.

wearing arbitrary clothing from only video streams is still not possible. Time-varying scene representations can also be reconstructed by means of shape-from-silhouette approaches [7], or with combined silhouette- and stereo-based methods [8]. Unfortunately, the measured models often lack detail if only a small number of input camera views is available and it is hard to preserve topological correspondences over time. Researchers have also used physics-based methods to track simple human motions if a kinematic skeleton is available [9]. However, the methods can not be directly applied to objects made of a variety of different materials, and they are not able to track arbitrarily dressed humans completely passively.

Instead, we present a robust skeleton-less approach to automatically capture the motion of a moving human subject and generate plausible and realistic surface deformations from multiple video streams without optical markers. Our algorithm is simple and versatile and enables us to directly animate a high quality static human scan from unaltered video footage which enables potential new applications in motion capture, computer animation and 3D Video.

The main contribution of this paper is a simple and robust method to automatically identify features on a moving human wearing everyday apparel, and track their 3D trajectories. It does not employ any a priori information about the subject, e.g. a kinematic skeleton, and can therefore be straightforwardly applied to other subjects, e.g. animals or mechanical objects. We also present a fast mesh deformation approach that uses only a handful of feature trajectories to directly and realistically animate a static human body scan making it performs the same motion as the captured subject. Our algorithm handles humans wearing arbitrary and sparsely textured clothing. As an additional benefit, it also preserves the mesh's connectivity over time.

The remainder of this paper is structured as follows: Sect. 2 reviews the most relevant related work and Sect. 3 briefly describes our overall framework. Thereafter, Sect. 4 details our automatic approach to identify features and track their 3D trajectories without optical markers. Sect. 5 describes our fast deformation scheme that is used to animate the static human model over the whole sequence according to the constraints derived from the estimated 3D point trajectories. Experiments and results with several captured real-world sequences are shown in Sect. 6, and the paper concludes in Sect.7.

2 Related Work

In our research we capitalize on ideas that have been published in the fields of object tracking, motion capture and scene reconstruction. For the sake of brevity, we refer the interested reader to overview articles on object tracking [10,11]. The following, is by no means a complete list of references from the other two research topics, but merely a summary of the most related categories of approaches.

Human motion is normally measured by marker-based or marker-less optical motion capture systems [1] that parameterize the data in terms of kinematic skeletons. Unfortunately, these approaches can not directly measure time-varying

body shape and they even fail to track people wearing loose apparel. To overcome this limitation, some methods use hundreds of optical markings [5] for deformation capture, combine a motion capture system with a range scanner [4,12] or a shape-from-silhouette approach [6], or jointly use a body and a cloth model to track the person [13]. Although achieving good results, most of these methods require active interference with the scene or require hand-crafted models for each individual.

Alternatively, shape-from-silhouette algorithms [7], multi-view stereo approaches [14], or methods combining silhouette and stereo constraints [8] can be used to reconstruct dynamic scene geometry. To obtain good quality results, however, several cameras are required and it is hard to generate connectivity-preserving dynamic mesh models.

Some passive methods extract 3D correspondences from images to track simple deformable objects [15] or cloth [16]. They can also be employed to jointly capture kinematic motion parameters and surface deformations of tightly dressed humans [17,18]. Researchers have also used physics-based shape models to track textiles [19,20] or simple articulated humans [9]. Unfortunately, none of these methods is able to track people dressed in arbitrary everyday apparel completely passively.

In contrast, we propose a skeleton-less method to directly capture the poses of a moving human subject and generate plausible surface deformations from only a handful of input video streams. This is achieved by first robustly identifying and tracking image features in 3D space. Thereafter, using the 3D trajectories of the features as constraints in a Laplacian mesh editing setting [21], the human model is realistically animated over time. By relying on differential coordinates, plausible shape deformations for the human scan are computed without having to specify explicit material parameters. Our algorithm is simple, robust, easy to implement and works even for moving subjects wearing wide and loose apparel.

3 Overview

An overview of our approach is shown in Fig. 1. Our system expects as input a multi-view video (MVV) sequence that shows the person moving arbitrarily. After acquiring the sequence, silhouette images are calculated via color-based background subtraction and we use the synchronized video streams to extract and track features in 3D space over time.

Our hybrid 3D point tracking framework jointly uses two techniques to estimate the 3D trajectories of the features from unmodified multi-view video streams. First, features in the images are identified using the Scale Invariant Feature Transform (SIFT). Furthermore, SIFT is able to match a feature to its corresponding one from a different camera viewpoint. This allows us to generate a set of pairwise pixel correspondences between different camera views for each time step of input video. Unfortunately, tracking the features over time using only local descriptors is not robust if the human subject is wearing sparsely textured clothing. Therefore, we use a robust dense optical flow method as an additional step to track the features for each camera view separately to fill the

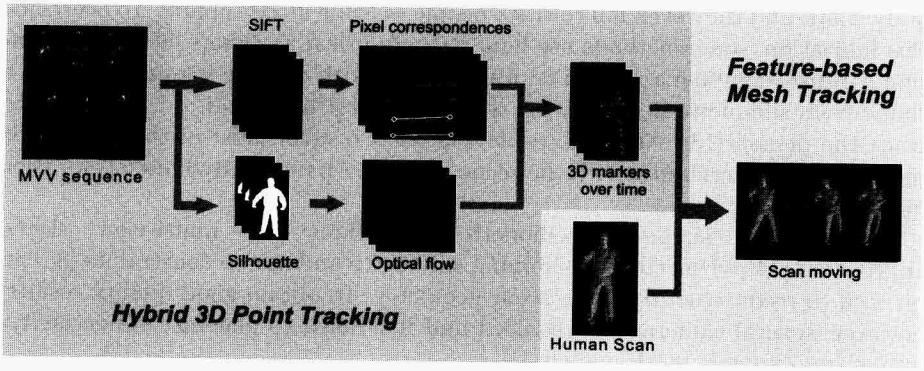

Fig. 1. Overview of our framework: given a multi-view video sequence showing a human performing, our method automatically identifies features and tracks their 3D trajectories. By applying the captured trajectories to a static laser-scan of the subject we are able to realistically animate a human model making it move the same as its real-world counterpart in the video streams.

gaps in the SIFT tracking. By merging both source of information we are able to reconstruct the 3D trajectories for many features over the whole sequence.

Our hybrid technique is able to correctly identify and track many 3D points. In addition to the estimation of 3D point positions, our approach also calculates a confidence value for each estimation. Using confidence-weighted feature trajectories as deformation constraints, our system robustly brings a static laser-scanned triangle mesh M of the subject into life by making it follow the motion of the actor recorded in the video frames.

4 Hybrid 3D Point Tracking

Our hybrid framework jointly employs local descriptors and dense optical flow to identify features and estimate their 3D positions over time from multiple calibrated camera views. In contrast to many other approaches [22,23,24], we developed an automatic tracking algorithm that works directly on the images without any a priori knowledge about the moving subject. It is our goal to create a simple and generic algorithm that can be used to track features on rigid bodies, articulated objects and non-rigidly deforming subjects in the same way.

The input to our algorithm comprises of synchronized video streams recorded from K cameras, each containing N video frames (Fig. 2a). In the first step, we automatically identify L important features, also called keypoints, for each camera view k and time step t and generate a set of local descriptors $F_{k,t} = \{f^0_{k,t}, \ldots, f^L_{k,t}\}$ using SIFT [25], Fig. 2b. We extract these features using the interest point detector proposed by Lowe [26] that is based on local 3D extrema in the scale-space pyramid built with difference-of-Gaussian filters. The local descriptors are built as a distinctive representation of the feature in an image from a patch of pixels in its neighborhood.

Since the SIFT descriptors are invariant to image scale, rotation, change in viewpoints, and change in illumination, they can be used to find corresponding features across different camera views. Given an image $I_{k,t}$, from camera view k and time step t, and the respective set of SIFT descriptors $F_{k,t}$, we try to match each element of $F_{k,t}$ with the set of keypoints from all other camera views. We use a matching function similar to [25], which assigns a match between $f^i_{k,t}$ and a keypoint in $F_{j,t}$ if the Euclidean distance between their invariant descriptor vectors is minimum. In order to discard false correspondences, nearest neighbor distance ratio matching is used with a threshold T_{MATCH} [27] .

After matching the keypoints across all K camera views at individual time steps, we gather all R correct pairwise matches into a list of pixel correspondences $C_t = \{c^0_t, \ldots, c^R_t\}$ by using all reliable matches found for each time step t (Fig. 2c). Each element $c^r_t = ((cam_u, P^i_t), (cam_v, P^j_t))$ stores the information about a correspondence between two different camera views, i.e. that pixel P^i_t in camera cam_u corresponds to pixel P^j_t in camera view cam_v at time t.

Unfortunately, tracking the features over time using only the list of correspondences C and connecting their elements at different time steps is not robust, because it is very unlikely that the same feature will be found at all time instants. This is specially true if the captured images show subjects performing fast movements, where features can be occluded for a long period of time, or when the subject wears everyday apparel with sparse texture. In the latter case, SIFT only detects a small number of keypoints per time step, which is usually not enough for tracking articulated objects. Therefore, in order to robustly reconstruct the 3D trajectories for the features we decided to use optical flow to track both elements of all c^r_t for each camera view separately, i.e. the pixel P^i_t is tracked using camera view cam_u and the pixel P^j_t using camera view cam_v.

The 2D flow-based tracking method works as follows: for each camera view k, we track all pixels over time using the warping-based method for dense optical flow proposed by Brox et al. [28]. After calculating the optical flow $\boldsymbol{o^t_k}(I_{k,t}, I_{k,t+1})$ between time step t and $t+1$ for camera k, we use $\boldsymbol{o^t_k}$ to warp the image $I_{k,t}$ and we verify for each pixel in the warped image if it matches the corresponding pixel in $I_{k,t+1}$. We eliminate the pixels that do not have a partner in $t+1$ and the pixels that belong to the background by comparing the warped pixels with the pre-computed silhouette $SIL_{k,t+1}$. This process is repeated for all consecutive time steps and for all camera views. As a result, we construct a tracking list $D_k = \{E^0, \ldots, E^g\}$ with G pixel trajectories for each camera view k (Fig. 2d). Each element $E^i = \{P^i_0, \ldots, P^i_N\}$ contains the positions of the pixel P^i_t for all time steps t.

The last step of our hybrid tracking scheme merges the optical flow tracking information with the list of correspondences to reconstruct the 3D trajectories for all features. We take pixel correspondences from all time steps into account. For instance, if a matching c^r_t is detected by SIFT only at the end of the sequence we are still able to recover the anterior positions of the feature by using the optical flow information.

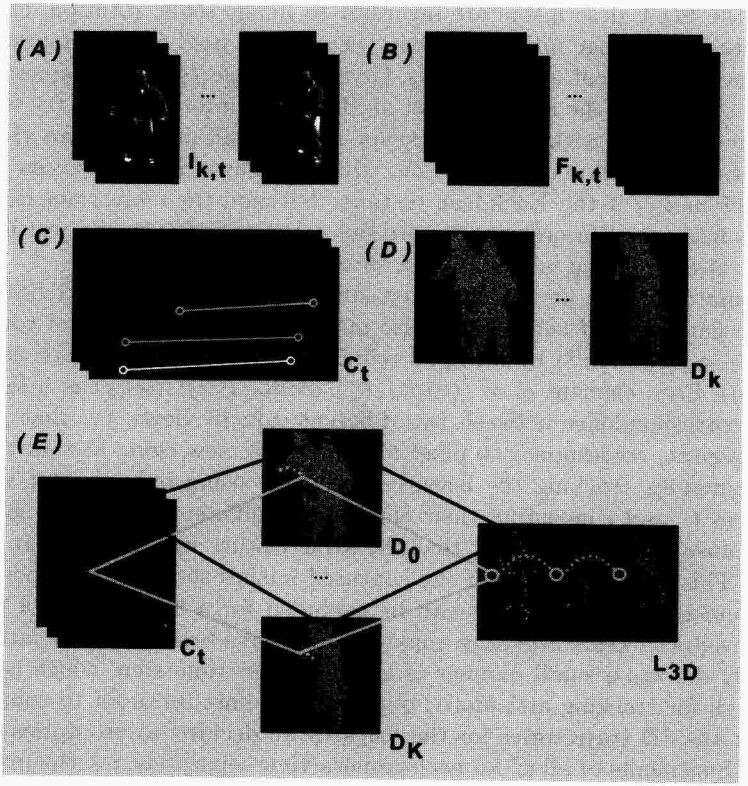

Fig. 2. Using the synchronized video streams as input (A), our hybrid approach first identifies features in the images using SIFT (B) and then matches these features between different pairs of camera views based on their descriptors (C). In addition, we track these features for each camera view separately using optical flow (D). At the end, reliable 3D trajectories for the features are reconstructed by merging both information (E).

For each entry $c_t^r = ((cam_u, P_t^i), (cam_v, P_t^j))$, we verify if the pixel P_t^i is found in D_{cam_u} and if the pixel P_t^j is found in D_{cam_v}. In case both elements are found, we estimate the position of the respective 3D point, $mm_r(t)$, for the whole sequence (Fig. 2e), otherwise c_t^r is discarded. The 3D positions are estimated by triangulating the viewing rays that start at the camera views cam_u and cam_v and pass through the respective image plane pixel at P_t^i and P_t^j. However, due to inaccuracies, these rays will not intersect exactly at a single point. However, we can compute a pseudo-intersection point $pos_t^r = \{x, y, z\}$ that minimizes the sum of squared distance to each pointing ray. We also use the inverse of this distance, cv_r, as a confidence measure indicating how reliable a particular feature has been located. If cv_r is below a threshold T_{CONF} we discard it, since it indicates that c_t^r assigns a wrong pixel correspondence between two different camera views.