Human Shape and Motion from Video

Pascal Fua
Computer Vision Laboratory, EPFL
CH-1015 Lausanne, Switzerland

Abstract

In recent years, because cameras have become inexpensive and ever more prevalent, there has been increasing interest in modeling human shape and motion from image data. This type of modeling has many applications, such as electronic publishing, entertainment, sports medicine and athletic training. This, however, is an inherently difficult task, both because the body is very complex and because the data that can be extracted from images is often incomplete, noisy and ambiguous. EPFL's Computer Vision Laboratory seeks to overcome these difficulties by using facial and body animation models, not only to represent the data, but also to guide the fitting process, thereby substantially improving performance. Start from sophisticated 3-D animation models, we reformulate them so that they can be used for data analysis in the three following research areas.

1 Augmented reality and 3-D tracking

In augmented reality applications, tracking and registration of cameras and objects are required because, to combine real and rendered scenes, we must project synthetic models at the right location in real images. As shown in Fig. 1, we have developed robust real-time methods for 3-D tracking of rigid objects and human faces [9, 10]. We formulate the tracking problem in terms of local bundle adjustment and merge the information from preceding frames with that provided by a very limited number of keyframes created during a training stage, which results in a real-time tracker that does not jitter or drift and can deal with significant aspect changes. We have also developed the fast 3-D object detection and pose estimation method [4, 5] which can be used to initialize or reinitialize the tracker in real-time. It relies on matching keypoints but, by contrast with previous methods that rely either on using ad hoc local descriptors or on estimating local affine deformations, the wide baseline matching of these keypoints is treated as a classification problem, in which each class corresponds to the set of all possible views of such a point. We synthesize a large number of views of individual keypoints of the object and train a classifier to recognize them. At run-time, we rely on this description to decide to which class, if any, an observed feature belongs. This formulation allows us to use powerful and fast classification methods to reduce matching error rates.

2 Face modeling from uncalibrated video sequences

In recent years, the movie industry has produced such realistic 3–D face models from images that we have come to take them for granted. However, a quick look at the credits

BMVC 2004 doi:10.5244/C.18.1

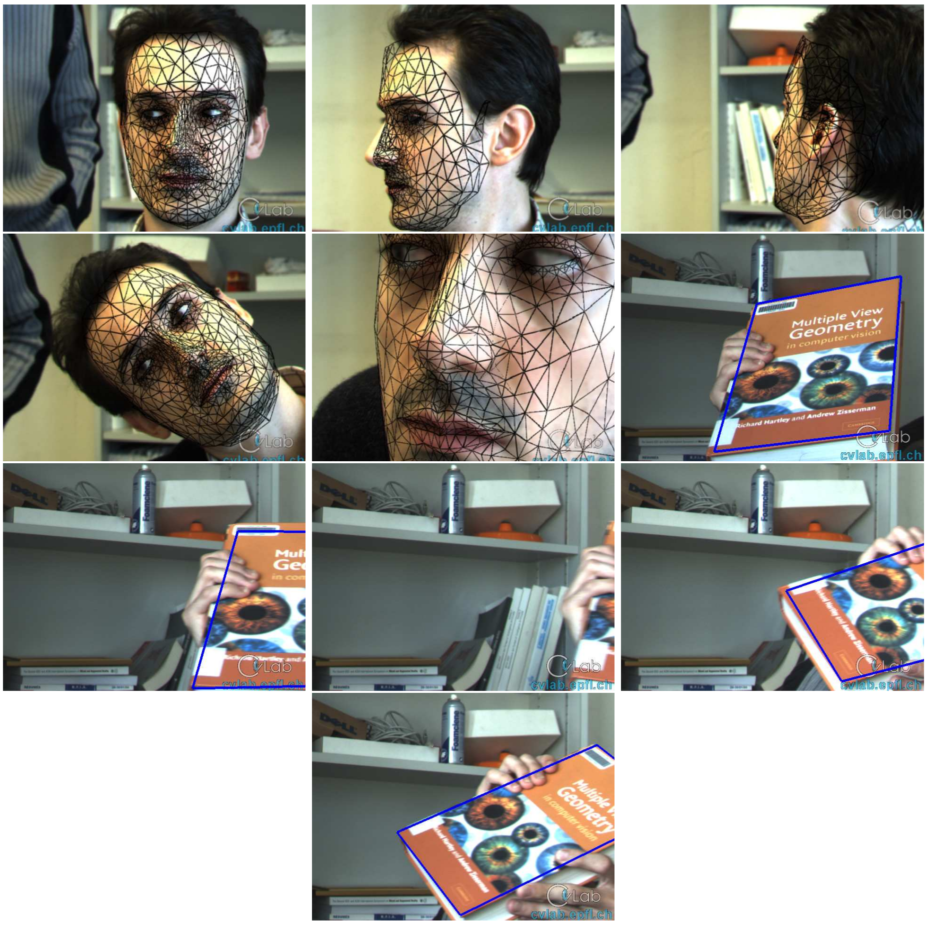
Multiple View Geometry in computer vision
Richard Hartley and Andrew Zisserman
cvlab.epfl.ch

Figure 1: The face is tracked in 3–D using a generic model that has not been tailored for this specific man. Note that the algorithm is robust to severe viewpoint changes and partial occlusions. The book is detected independently in each frame. Both algorithms run at 25Hz on a standard PC.

at the end of a movie such as “The Matrix Reloaded” and at the budgets that are involved, should alert the careful scientist to the fact that this is a misperception. To obtain similar results at much lower cost, it therefore still makes sense to address the structure-from-motion problem in the context of head modelling from video sequences for which calibration data is not available. Fig. 2 depicts our most recent approach [2] which relies on optimizing the shape parameters of a sophisticated PCA based model [1] given pairwise image correspondences. All that is required is enough relative motion between camera and subject so that we can derive structure from motion. By matching the results against laser scanning data, we have shown that its precision is excellent and can be predicted as a function of the number and quality of the correspondences. This is important to establish the appropriate compromise between processing speed and quality of the results.

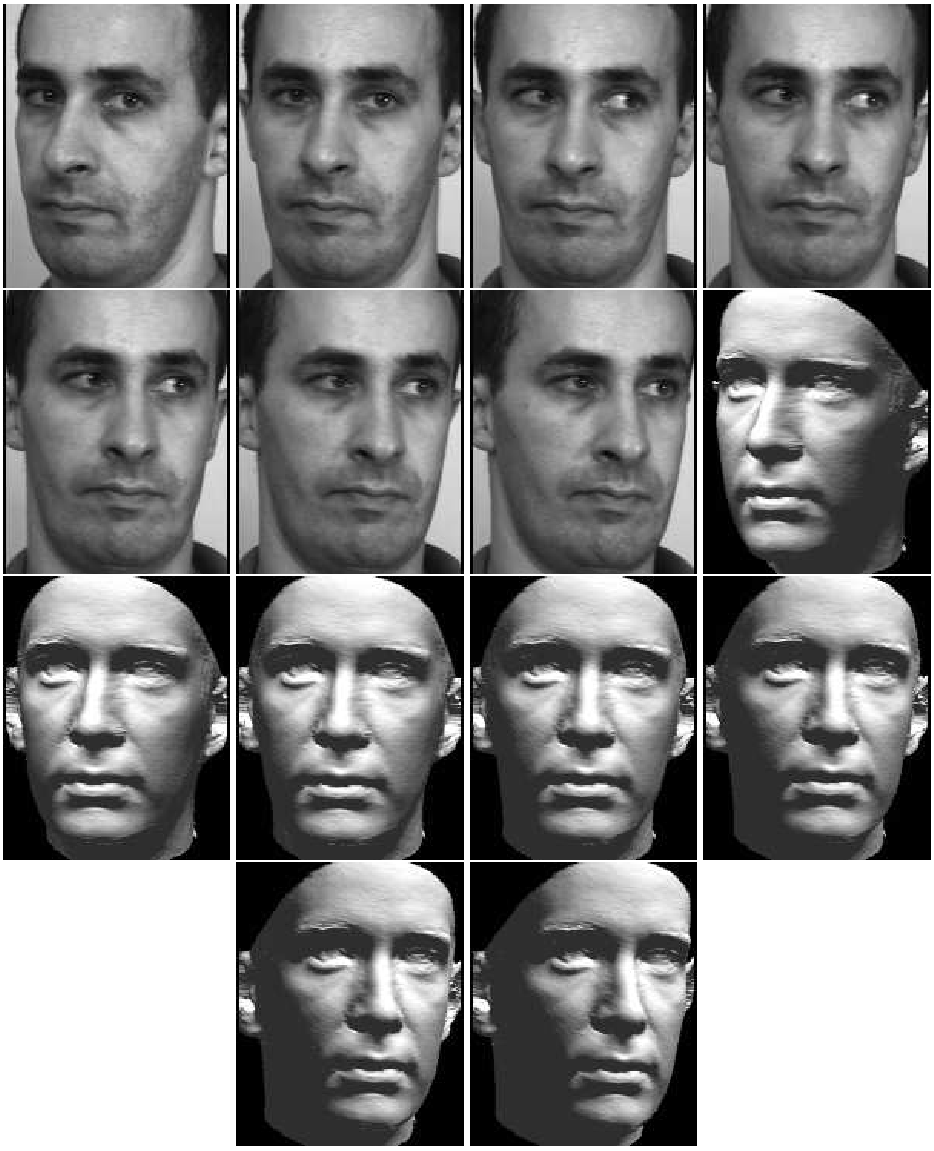

Figure 2: A short 7 image video sequence. Shaded reconstruction using all 7 images. The median deviation from a laser-scan of the same face is smaller 0.5mm.

3 Motion capture from video sequences

We have developed a framework for 3–D shape and motion recovery of articulated deformable objects. We proposed a formalism that incorporates the use of implicit surfaces into earlier robotics approaches that were designed to handle articulated structures [6] and impose hierarchical joint limits [3]. We have further increased its robustness by introducing temporal motion models based on Principal Component Analysis to formulate the tracking problem as one of minimizing differentiable objective functions [8]. Our experiments, such as the one depicted by Fig. 3, show that the differential structure of these objective functions is rich enough to take advantage of standard deterministic optimization methods, whose computational requirements are much smaller than those of probabilistic ones and can nevertheless yield very good results even in difficult situations. Further work will focus both on further refining our models to improve their ability to constrain the interpretation of noisy image-data and on increasing the quality of this data [7].

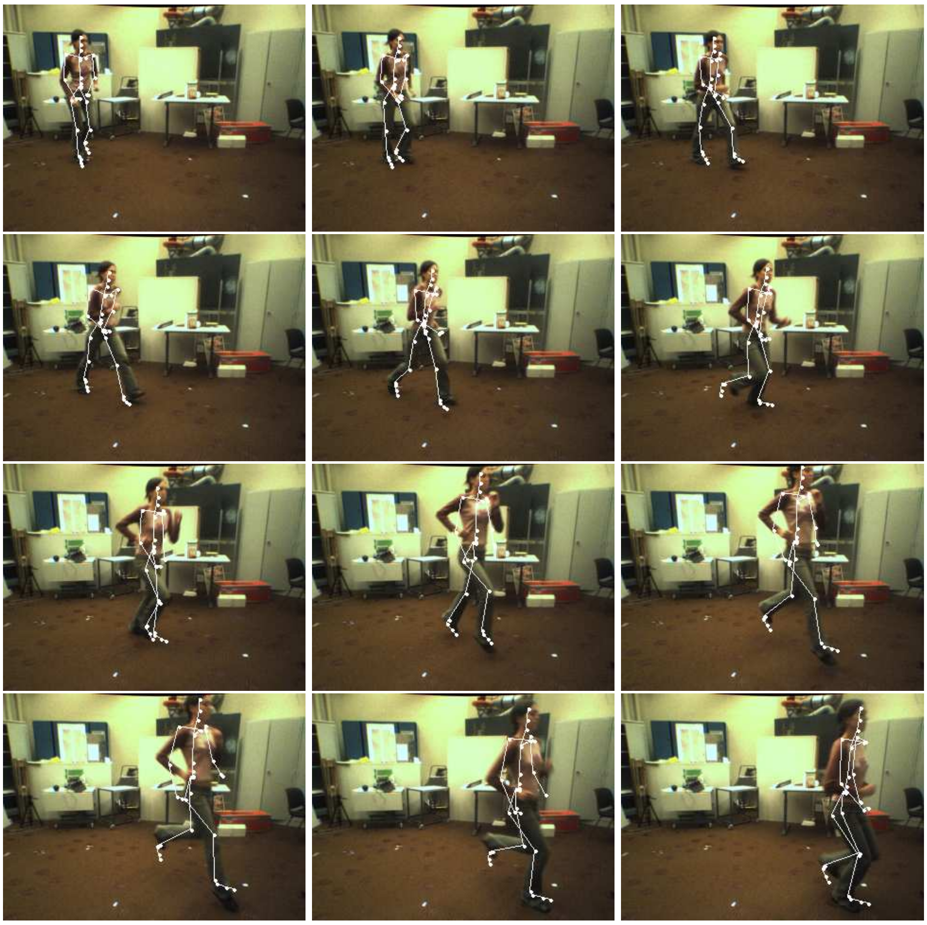

Figure 3: Tracking results for a transition from walking to running. In the first four frame the subject is running, then the following two or three frames are the transition. The sequence end with a running.

References

[1] V. Blanz and T. Vetter. A Morphable Model for The Synthesis of 3–D Faces. In *Computer Graphics, SIGGRAPH Proceedings*, Los Angeles, CA, August 1999.

[2] M. Dimitrijevic, S. Ilic, and P. Fua. Accurate Face Models from Uncalibrated and Ill-Lit Video Sequences. In *Conference on Computer Vision and Pattern Recognition*, Washington, DC, June 2004.

[3] L. Herda, R. Urtasun, and P. Fua. Hierarchical Implicit Surface Joint Limits to Constrain Video-Based Motion Capture. In *European Conference on Computer Vision*, Prague, Czech Republic, May 2004.

[4] V. Lepetit, J. Pilet, and P. Fua. Point Matching as a Classification Problem for Fast and Robust Object Pose Estimation. In *CVPR*, Washington, DC, June 2004.

[5] V. Lepetit, J. Pilet, and P. Fua. Towards Recognizing Feature Points using Classification Trees. Technical report, EPFL, 2004.

[6] R. Plänkers and P. Fua. Articulated Soft Objects for Multi-View Shape and Motion Capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003.

[7] A. Shahrokni, T. Drummond, and P. Fua. Markov-based Silhouette Extraction for Three–Dimensional Body Tracking in Presence of Cluttered Background. In *These proceedings*, 2004.

[8] R. Urtasun and P. Fua. 3D Human Body Tracking using Deterministic Temporal Motion Models. In *European Conference on Computer Vision*, Prague, Czech Republic, May 2004.

[9] L. Vacchetti, V. Lepetit, and P. Fua. Fusing Online and Offline Information for Stable 3–D Tracking in Real-Time. In *Conference on Computer Vision and Pattern Recognition*, Madison, WI, June 2003.

[10] L. Vacchetti, V. Lepetit, and P. Fua. Stable real-time 3d tracking using online and offline information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004. In press.