Computers & Graphics xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on Motion in Games 2016

Real-time labeling of non-rigid motion capture marker sets

Q1 Simon Alexanderson [a,*], Carol O'Sullivan [b], Jonas Beskow [a]

[a] KTH - Speech, Music and Hearing, Lindsetdtsv. 24, Stockholm 10044, Sweden
[b] Trinity College Dublin, College Green, Dublin 1, Ireland

ARTICLE INFO

Article history:
Received 1 August 2017
Revised 1 October 2017
Accepted 2 October 2017
Available online xxx

Keywords:
Animation
Motion capture
Hand capture
Labeling

ABSTRACT

Passive optical motion capture is one of the predominant technologies for capturing high fidelity human motion, and is a workhorse in a large number of areas such as bio-mechanics, film and video games. While most state-of-the-art systems can automatically identify and track markers on the larger parts of the human body, the markers attached to the fingers and face provide unique challenges and usually require extensive manual cleanup. In this work we present a robust online method for identification and tracking of passive motion capture markers attached to non-rigid structures. The method is especially suited for large capture volumes and sparse marker sets. Once trained, our system can automatically initialize and track the markers, and the subject may exit and enter the capture volume at will. By using multiple assignment hypotheses and soft decisions, it can robustly recover from a difficult situation with many simultaneous occlusions and false observations (ghost markers). In three experiments, we evaluate the method for labeling a variety of marker configurations for finger and facial capture. We also compare the results with two of the most widely used motion capture platforms: Motion Analysis Cortex and Vicon Blade. The results show that our method is better at attaining correct marker labels and is especially beneficial for real-time applications.

1. Introduction

Optical marker-based motion capture is a mature and dominant technology for capturing detailed human motion in many areas such as bio-mechanics, film and video games. The technology provides many desirable features such as high accuracy and sampling rates and can be used as a single means to capture body and finger motion as well as facial expression. Among the main challenges for optical motion capture using passive markers is the identification and tracking of the markers, commonly referred to as labeling. The difficulties arise due to the fact that the markers look identical from the point of view of the system and their identities need to be inferred from structural cues or tracked over time, something that is further challenging in cases of severe occlusions.

Current state-of-the-art motion capture systems can reliably label markers on the larger parts of the human body, also in large capture volumes. However, markers on the more articulated body parts, such as the face and fingers, pose unique challenges and usually require extensive manual labelling. For facial capture, alternative markerless methods (such as video based tracking using head-mounted cameras) has gained in popularity, but this adds cost and complexity to the setup, and there are still many domains in which head-mounted cameras are too intrusive to be used. For finger capturing, the only viable solutions for large volumes are to use either data gloves or sparse marker sets with optical motion capture, [1]. In our work, we connect to the recent advances in data-driven methods to produce high quality hand and finger animation from sparse marker sets [2–4], and address the problems of automatic labelling of such markers. Sparse marker sets prove to be especially challenging for existing labeling algorithms. This is mainly due to the fact that sparsity reduces the structural information available to the point where underlying skeleton models, commonly used in existing labeling algorithms, are difficult to apply.

In this paper, we present an extended version of our paper on robust algorithms for automatic labeling of finger markers, [5]. In addition to previously reported work, we show how our method can be extended to simultaneously label multiple marker sets in close interaction, and present new results of labeling face and finger markers in a full performance capture setup. We also show how our method integrates with data-driven methods for reconstructing full marker sets from sparse data, and hence allows users to reduce the number of markers in a capture without significant loss of quality.

At the core of our system is an algorithm to generate multiple assignment hypotheses based on the spatial distribution of

* Corresponding author.
E-mail address: simonal@kth.se (S. Alexanderson).

https://doi.org/10.1016/j.cag.2017.10.001
0097-8493/© 2017 Published by Elsevier Ltd.

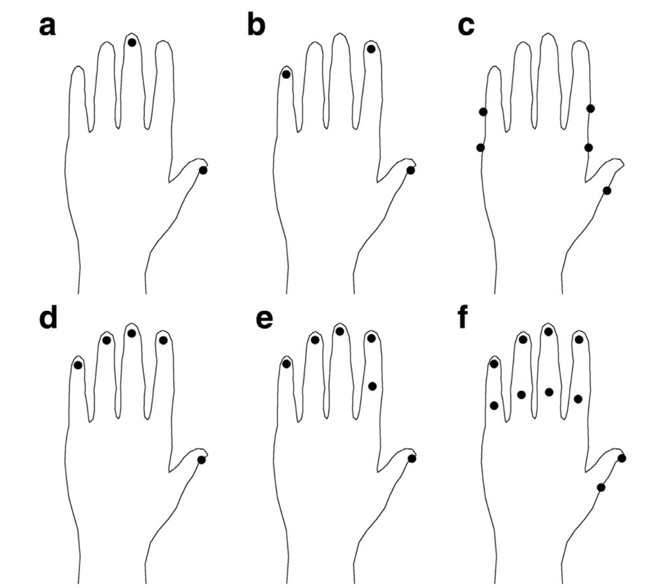

Fig. 1. A selection of sparse marker sets for finger capture: (a) and (b) [6]; (c) Optitrack Motive; (d) [7,8]; (e) [3]; and (f) [9]. The top row shows common marker sets used in the industry, and the bottom row shows the recommended marker sets from the research community. Note the large marker separation in the top row, facilitating automatic labeling.

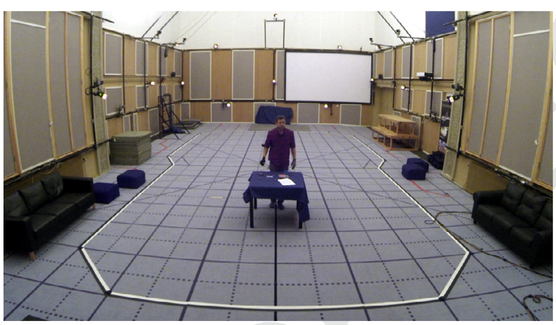

Fig. 2. Capture volume of 7 m × 12 m × 5 m.

the markers, and another algorithm to select the best sequence of assignments in time. A key characteristic of our method is the domain in which the assignment hypotheses are generated. While other methods generate assignments from the *temporal* domain, i.e. from the predicted marker positions at each frame, and use an initialization phase (usually involving a T-pose) to commence tracking, our method continuously generates a fixed set of assignment hypotheses from the *spatial* domain, and treats tracking as an optimization problem to find the most probable path through the hypothesis space. In this way, our method can continuously reinitialize the marker labels even after long occlusions. By using multiple assignment hypotheses, no hard decisions are made at times where the assignments are ambiguous due to occlusions and/or ghost markers, and the algorithm has a chance to correct errors as more evidence becomes available.

We evaluate our method in three experiments. The first experiment covers finger capturing using a variety of different marker sets described in the literature (see Fig. 1), and shows that our method is able to provide correct labels for over 99.6% of the data for all of the marker sets. The second experiment covers finger capturing in a large volume (see Fig. 2). Bench-marked against two of the most dominant commercial platforms, Motion Analysis Cortex[1], and Vicon Blade[2], our method is better at attaining correct marker labels in general and is particularly beneficial for fragmented data. The third experiment covers simultaneous labelling of face and finger markers in a full performance capture and demonstrates how the method is used in conjunction with data-driven methods to generate rich data sets from sparse markers.

As our method is working in real-time, it is of special use to the video-games and film industries, which require large capture volumes for in-game motion and cinematics, and real-time capabilities for Virtual Reality, Previs and Virtual Production.

2. Related work

Early marker labeling techniques emerge from the field of Multiple Target Tracking (MTT) [10], which was originally developed for tracking radar plots. One of the most successful MTT algorithms is Multiple Hypothesis Tracking [11], which allows for soft decision making when the observations are noisy and the tracking situation is ambiguous. A limitation of using MTT algorithms for motion capture is that they do not take structural information into account, and thus needs to be manually initialized at the first time frame as well as after longer periods of gaps. In most motion capture scenarios, the motions of the markers are correlated in some way, which may be exploited for labeling. Gennari et al. [12] integrate shape constraints in MTT, but do not initialize marker identities or use multiple hypotheses. Also Yu et al. [13] exploit structural information, but their algorithm requires a large number of markers and is not suitable for sparse, non-rigid marker sets.

Other studies focus on simultaneous labeling and skeleton solving using an underlying skeleton model. Ringer and Lasenby [14] developed a multiple hypotheses tracker and demonstrate their method on human body motion. Meyer et al. [15] used a probabilistic framework for automatic online labeling of full-body marker sets, and Schubert et al. [16] extend this method to be able to initialize the tracking using an arbitrary pose. As opposed to our approach, these methods require dense enough marker sets to uniquely define the pose of the underlying skeleton model. Our method is developed for sparse marker sets and data-driven pose estimation, where as few as 3 markers may be used to drive more than 20 degrees of freedom of finger motion. Recently, Maycock et al. [9] developed a labeling system using an inverse kinematics (IK) based skeleton, and demonstrated it for capturing hand and finger motion. However, their method requires a specialized initialization pose and does not use multiple hypotheses, and it is not clear how it would reinitialize in cases where several markers are occluded for longer time periods. In a study by Akhter et al. [17], a spatiotemporal model was developed to perform simultaneous labeling and gap-filling. The method was demonstrated on a dense set of 315 facial markers. However, in contrast to our domain where only a few loosely correlated markers exist, their data set contains a large amount of spatiotemporal correlation, making it possible to deduce lost marker positions from the trained model.

The capturing of hand motion is an active research field with many recent publications (see the state-of-the-art report [1] for an overview). While there have been major improvement in markerless methods based on computer vision techniques and depth sensors, these methods still impose severe restrictions, e.g. on capture volumes, frame rates and tracking of parts that are in physical contact. According to [1], they are only appropriate in small volumes and have difficulties in reconstructing complex hand shapes. Other techniques exist based on instrumented gloves such as the

[1] http://www.motionanalysis.com.
[2] http://www.vicon.com.