Chapter 5
KSM Based Machine Learning for Markerless Motion Capture

Therdsak Tangkuampien
Monash University, Australia

David Suter
Monash University, Australia

ABSTRACT

A marker-less motion capture system, based on machine learning, is proposed and tested. Pose information is inferred from images captured from multiple (as few as two) synchronized cameras. The central concept of which, we call: Kernel Subspace Mapping (KSM). The images-to-pose learning could be done with large numbers of images of a large variety of people (and with the ground truth poses accurately known). Of course, obtaining the ground-truth poses could be problematic. Here we choose to use synthetic data (both for learning and for, at least some of, testing). The system needs to generalize well to novel inputs: unseen poses (not in the training database) and unseen actors. For the learning we use a generic and relatively low fidelity computer graphic model and for testing we sometimes use a more accurate model (made to resemble the first author). What makes machine learning viable for human motion capture is that a high percentage of human motion is coordinated. Indeed, it is now relatively well known that there is large redundancy in the set of possible images of a human (these images form some sort of relatively smooth lower dimensional manifold in the huge dimensional space of all possible images) and in the set of pose angles (again, a low dimensional and smooth sub-manifold of the moderately high dimensional space of all possible joint angles). KSM, is based on the KPCA (Kernel PCA) algorithm, which is costly. We show that the Greedy Kernel PCA (GKPCA) algorithm can be used to speed up KSM, with relatively minor modifications. At the core, then, is two KPCA’s (or two GKPCA’s) - one for the learning of pose manifold and one for the learning image manifold. Then we use a modification of Local Linear Embedding (LLE) to bridge between pose and image manifolds.

DOI: 10.4018/978-1-60566-900-7.ch005
1. OVERVIEW

Humans can look at an image of a human and “see” what pose they are in (i.e., can infer, to some accuracy, the joint angles). We have built a system to do something similar: pose information is inferred from images captured from multiple (as few as two) synchronized cameras. The central concept of which, we call: Kernel Subspace Mapping (KSM) (Section 5).

The concept can be seen in Figure 1. The image (we extract silhouettes) to pose mapping is to be learnt (left hand side) and then used (right hand side). The learning could be done with large numbers of images of a large variety of people (and with the ground truth poses accurately known). Of course, obtaining the ground-truth poses could be problematic: In principle, this could be done using a commercial (expensive) motion capture system but we choose to use synthetic data (both for learning and for, at least some of, testing). Of course, the system needs to generalize well to novel inputs. That is unseen poses (not in the training database) and unseen actors. For the learning we use a generic and relatively low fidelity computer graphic model (left of the figure) and for testing we sometimes use a more accurate model (made to resemble the first author).

What makes machine learning viable for human motion capture is that a high percentage of human motion is coordinated [Safonova et al., 2004; Char and Hodgins, 2005]. Indeed, it is now relatively well known that there is large redundancy in the set of possible images of a human (these images form some sort of relatively smooth lower dimensional manifold in the huge dimensional space of all possible images) and in the set of pose angles (again, a low-dimensional and smooth sub-manifold of the moderately high dimensional space of all possible joint angles). Figure 7 shows these subspaces.

There have been many experiments on the application of learning algorithms (such as Principal Components Analysis (PCA) [Jolliffe, 1986; Smith, 2002] and Locally Linear Embedding (LLE) [Saul and Roweis, 2000]) in learning low

Figure 1. Diagram to summarize the training and testing process of Kernel Subspace Mapping, which learns the mapping from image to the normalized Relative Joint Centers (RJC) Pose space (section 1.3.1). Note that different mesh models are used in training and testing. The generic model [top left] is used to generate training images, whereas the accurate mesh model of the author (Appendix 3) is used to generate synthetic test images.