This thesis consists of three parts. In the first part (Chapters 1-2), we introduce the overview, motivation, and contribution of our work, and we extensively survey the current literature for 6 topics that are closely related to our work. In the second part (Chapters 3-6), we thoroughly explore the concept of `Self-Similarity" in two challenging scenarios, namely, the Action Recognition and the Motion Retrieval. We build three-dimensional volume representations for both scenarios, and devise effective techniques that can produce very compact representations, which can successfully encode the internal dynamics of data. In the third part (Chapter 7), we explore the challenging action spotting problem, and propose a feature-independent unsupervised framework that is very effective in spotting action under various real situations, even under heavily perturbed conditions.

For action recognition, we introduce a generic method that does not depend on one particular type of input feature vector. We make three main contributions: (i) We introduce the concept of Joint Self-Similarity Volume (Joint SSV) for modeling dynamical systems, and show that by using a new optimized rank-1 tensor approximation of Joint SSV one can obtain compact low-dimensional descriptors that very accurately preserve the dynamics of the original system, e.g. an action video sequence; (ii) The descriptor vectors derived from the optimized rank-1 approximation make it possible to recognize actions without explicitly aligning the action sequences of varying speed of execution or difference frame rates; (iii) The method is generic and can be applied using different low-level features such as silhouettes, histogram of oriented gradients (HOG), etc. Hence, it does not necessarily require explicit tracking of features in the space-time volume. Our experimental results on five public datasets demonstrate that our method produces remarkably good results and outperforms all baseline methods.

For motion retrieval, we present a framework that allows for a flexible and an efficient retrieval of motion capture data in huge databases. The method first converts an action sequence into a self-similarity matrix (SSM), which is based on the notion of self-similarity. This conversion of the motion sequences into compact and low-rank subspace representations greatly reduces the spatiotemporal dimensionality of the sequences. The SSMs are then used to construct order-3 tensors, and we propose a low-rank decomposition scheme that allows for converting the motion sequence volumes into compact lower dimensional representations, without losing the nonlinear dynamics of the motion manifold. Thus, unlike existing linear dimensionality reduction methods that distort the motion manifold and lose very critical and discriminative components, the proposed method performs well, even when inter-class differences are small or intra-class differences are large. In addition, the method allows for an efficient retrieval and does not require the time-alignment of the motion sequences. We evaluate the performance of our retrieval framework on the CMU mocap dataset under two experimental settings, both demonstrating very good retrieval rates.

For action spotting, our framework does not depend on any specific feature (e.g. HOG/HOF, STIP, silhouette, bag-of-words, etc.), and requires no human localization, segmentation, or framewise tracking. This is achieved by treating the problem holistically as that of extracting the internal dynamics of video cuboids by modeling them in their natural form as multilinear tensors. To extract their internal dynamics, we devised a novel Two-Phase Decomposition (TP-Decomp) of a tensor that generates very compact and discriminative representations that are robust to even heavily perturbed data. Technically, a Rank-based Tensor Core Pyramid (Rank-TCP) descriptor is generated by combining multiple tensor cores under multiple ranks, allowing to represent video cuboids in a hierarchical tensor pyramid. The problem then reduces to a template matching problem, which is solved efficiently by using two boosting strategies: (1) to reduce search space, we filter the dense trajectory cloud extracted from the target video; (2) to boost the matching speed, we perform matching in an iterative coarse-to-fine manner. Experiments on 5 benchmarks show that our method outperforms current state-of-the-art under various challenging conditions. We also created a challenging dataset called Heavily Perturbed Video Arrays (HPVA) to validate the robustness of our framework under heavily perturbed situations.
Educational Career:
Bachelor's of Computer Science, BS, 2006, Central South University
Master's of Computer Science, MS, 2009, Central South University

Committee in Charge:
Hassan Foroosh, Chair, Electrical Engineering and Computer Science
Rahul Sukthankar, Google Research
Marshall Tappen, Amazon.com
Charles Hughes, Electrical Engineering and Computer Science
Mike Moshell, School of Visual Arts and Design (SVAD), University of Central Florida

Approved for distribution by Hassan Foroosh, Committee Chair, on January 1, 2016.

The public is welcome to attend.