### Abstract

This paper addresses the problem of human action recognition by introducing a sparse representation of image sequences as a collection of spatiotemporal events that are localized at points that are salient both in space and time. We detect the spatiotemporal salient points by measuring changes in the information content of pixel neighborhoods not only in space but also in time. We introduce an appropriate distance metric between two collections of spatiotemporal salient points that is based on the Chamfer distance and an iterative linear time warping technique that deals with time expansion or time compression issues. We propose a classification scheme that is based on Relevance Vector Machines and on the proposed distance measure. We present results on real image sequences from a small database depicting people performing 19 aerobic exercises.