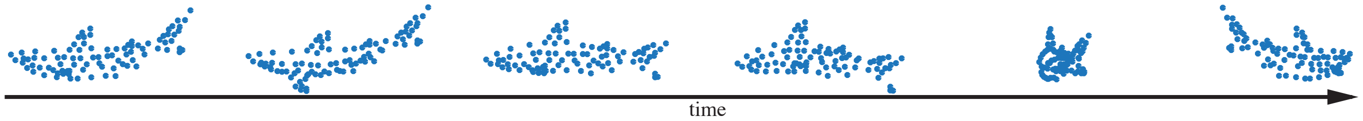
Formal Aspects of Non-Rigid-Shape-from-Motion Perception

Vicky Froyen & Qasim Zaidi



Our world is full of objects that deform over time, for example animals, trees and clouds. Some articulated objects are globally non-rigid but with rigid parts (e.g. birds, humans). For other objects, even their parts undergo non-rigid transformations (e.g. fish, faces). Rigorously defining a unique global motion for deforming objects presents a challenge, because an infinite set of pairs of rigid motions and shape deformations can describe the overall change. Yet, the human visual system seems to readily disentangle object motions from non-rigid deformations, in order to categorize objects, recognize the nature of actions such as running or jumping, and even to infer intentions.

A large body of experimental work has been devoted to extracting rigid structure from motion, but there is little experimental work on the perception of non-rigid 3-D shapes from motion (e.g. Jain, 2011). Similarly, until recently, almost all formal work had concentrated on the rigid case, e.g. Tomasi & Kanade's (1992) factorization approach, which shows how to decompose the image stream into a 3D shape matrix and an orthonormal rotation matrix. In the last fifteen years, Computer Vision researchers have made modeling advances in non-rigid structure from motion,. In one class of solutions, non-rigid 3D shapes are modeled as a weighted set of pre-learned basis shapes (e.g. Brand & Bhotika, 2001). Another class set out to solve the problem without any prior model as a generalization of the factorization approach (Bregler et al., 2000). Since for non-rigid shapes the 3D shape changes at each instant, it is defined as a linear combination of a small set of basis shapes, that are derived from an SVD of the image stream, sometimes regularized by priors of smoothness for shapes and deformations (Torresani et al., 2003). Based on the observation that deformations of natural objects are often cyclical, later approaches (e.g. Akhter et al., 2008) have approximated the 3D shape matrix as a linear combination of a small number of motion trajectory bases (the dual to the shape bases).

Both approaches have their inherent advantages and applications. With shape bases, the first basis gives an estimate of the average 3D shape, which could help with object recognition if it is general across motion streams. With trajectory bases, the weights provide efficient descriptors of the deformations that the objects are undergoing, and help with action recognition. Each of these dominant ideas has gone through many iterations of incremental improvement to handle missing data, noisy input, and perspective and orthogonal camera input. In this talk we will present the history of these advances, while examining the validity of the assumptions, and the performance of the models in the light of what we know about human vision.

1) No differentiation is made between camera/observer motion and object motion, whereas humans can distinguish the two situations (Warren & Rushton, 2009). 2) If derived basis shapes are affected by specific deformations during the image stream, they would not be useful for generalizing to future deformations. 3) Further, shape based models would give the same results to randomized image sequences, because they dont use motions per se, whereas human do need the motion information present. 4) Humans process motion fields by using prediction (e.g. Graf et al., 2005) and grouping (e.g. Johansson, 1950), but this has seldom been incorporated in computational models (Pentland & Horowitz, 1992; Brox & Malik, 2010). 5) Computational schemes rely heavily on a preprocessing step to extract features and track them reliably throughout an image sequence, without involving filters that process velocity patterns, as are found in primate cortex (Zhang, Sereno & Sereno, 1993). 6) Current models can't cope with severe cases of occlusion. 7) Trajectory bases estimated for slow to stationary cameras/observers can result in very bad 3D structural estimates (Park et al., 2010).

This talk will discuss how these models can be modified for human vision, particularly by testing assumptions in psychophysical and physiology experiments. We hope that this will spark interest in this new and exciting topic of research, and create cross talk between Computer Scientists and Vision Researchers.