Shape workshop – VSS 2011

9:00 Corrado Caudek: 3D shape perception and action in the peripersonal space
9:30 Sinisa Todorovic: Recognizing Human Activities by their Space-Time Shapes
10:00 Richard Murray: 3D shape perception: priors and contextual cues

10:30 Coffee Break

11:00 Song-Chun Zhu: Object Representation: Structure vs. Appearance and 3D vs. 2D - an Information Theoretical Perspective
11:30 Michael S. Lewicki: Perceptual organization of boundaries in natural scenes
12:00 Björn Ommer: Beyond the Sum of Parts: Voting by Grouping Dependent Image Fragments

12:30 Lunch

2:00 Ken Nakayama: Subjective Contours
2:30 David Jacobs: Understanding Shape by Comparing Images
3:00 Rudiger von der Heydt: Contour grouping in the visual cortex

3:30 Coffee Break

4:00 John Tsotsos: Attending to Shape
4:30 Edward H. Adelson: Surfaces, materials, and shape
5:00 Ben Kimia: Perceptual Fragments: Bottom-Up and Top-Down Use of Shape in Object Recognition

5:30 Panel Discussion

Abstracts

Corrado Caudek

3D shape perception and action in the peripersonal space

I will discuss the results of some recent investigations concerning the mechanisms underlying our perceptions of the 3D space and our interactions with the physical world. Our research is motivated by the following questions: Which kind of information is actually used by the visuomotor system? Is this information sufficient for recovering a veridical Euclidean representation of 3D space? Is it possible that metric information is not needed in most natural situations for the control of motor behavior? Are perception and action mediated by different systems? I will question the idea that the interactions between an agent and the physical world are necessarily mediated by an internal Euclidean metric representation. I will argue that this approach is certainly sensible from an engineering point of view, but it is not necessarily the best starting point for studying the functioning of a biological system. I will present empirical evidence indicating that active vision exhibits systematic biases that are similar to those found in passive vision. I will show that reaching and grasping movements, in absence of haptic feedback, exhibit systematic distortions that are similar to those found in our perception of the 3D layout and objects. I will interpret these findings as suggesting that (i) the visuomotor and perceptual systems mostly rely on unscaled 3D information directly specified by sensory data, (ii) the scaling of direct sensory information is obtained from experience, but it is not necessarily veridical, and (iii) in the absence of haptic feedback, reaching movements in the peripersonal space remain systematically distorted. In summary, I will discuss and provide support to the hypothesis that the visuomotor system relies heavily on recalibration and it is very plastic. Evidences of this plasticity will be presented here and in a related talk at the conference.

David Jacobs
Understanding Shape by Comparing Images

A fundamental problem in shape understanding is to determine whether two images depict shapes that come from the same object or class. This is difficult because the shape of an object can change due to deformations or articulations, and shape can differ considerably between objects from the same class. Moreover, the appearance of an object's shape can depend on viewpoint or lighting. One way to cope with this variability is by developing methods of image comparison that take account of these sources of change. This approach does not require us to reconstruct the shape of an object from an image, which may be a difficult or even an underconstrained problem. However, it does require us to create new algorithms that compare images using an understanding of how changes in shape, viewpoint and lighting can affect appearance. I will discuss a number of such methods, and highlight problems in image comparison that I still find mysterious and intriguing.

Benjamin Kimia
Perceptual Fragments: Bottom-Up and Top-Down Use of Shape in Object Recognition

The bottom-up "segmentation followed by recognition" strategy has for some time now given way to feature-based discriminative recognition with significant success. As the number of categories and exemplars per category increases, however, low-level features are no longer sufficiently discriminative, motivating the construction and use of more complex features. It is argued here that these complex features will necessarily be encoding shape and this in turn requires curves and regions, thus reviving aspects of bottom-up segmentation strategies. We suggest that the demise of segmentation was due to prematurely committing to a grouping in face of ambiguities and propose a framework for representing multiple grouping options in a containment graph. Specifically, we use contour symmetry to partition the image into atomic fragments and define transforms to iteratively grow these atomic fragments into mote distinctive perceptual fragments, the nodes of the containment graph. We also briefly present a fragment-based language for generating shapes and the use of fragments in top-down category recognition. The bottom-up and top-down processes are then integrated by interaction through the mid-level representation of perceptual fragments.

Michael Lewicki
Perceptual organization of boundaries in natural scenes

We readily perceive contours and surfaces in complex natural scenes. At the level of simple visual features, however, these more abstract structures are difficult to extract, because image patterns of both boundaries and surface regions are highly variable. What then are the computations that can deduce intrinsic structure from the raw sensory variability? In this talk, I will discuss an approach that is based on learning statistical distributions of local regions in a visual scene. This approach generalizes the theory of efficient coding for learning image features to hierarchical models. The central hypothesis is learning these local distributions allows the visual system to generalize across similar local image regions, i.e. textures within a surface or texture boundaries along a contour. Joint activity in the model encodes the probability distribution over their inputs and forms stable representations across complex patterns of variation. In addition, units in the model exhibit a diverse range of non-linear properties observed in neurons in visual cortex and provide a novel functional explanation for their role in visual perception.

This is joint work with Yan Karklin and Chris DiMattina.

Richard Murray
3D shape perception: priors and contextual cues

Image shading is a fundamental but highly ambiguous shape cue that results from the interaction of illumination, surface geometry, and material properties. This deep ambiguity means that any attempt to use shading as a shape cue must rely on strong assumptions about what lighting conditions and surface configurations are most probable. This raises two questions. First, what assumptions does human vision make about lighting and surfaces? Second, how are these assumptions (e.g., the light-from-above prior) reconciled with cues in specific scenes (e.g., lighting direction cues), given that they will often be in conflict? I will describe recent work on measuring statistical properties of natural illumination, and discuss evidence that regularities in natural illumination revealed by these studies guide human perception of surface properties like shape and reflectance. I will also describe work demonstrating that a cue combination strategy based on circular statistics guides the reconciliation between lighting priors and lighting cues.

Ken Nakayama
Subjective Contours

The literature on subjective contours is voluminous with a very long history. The demonstrations are some of the most powerful in the study of vision. Yet, with all this activity, there is little consensus as to how subjective contours should be interpreted and understood. In this talk I will present what I think are some of the most significant demonstrations that remain of theoretical relevance today.

Björn Ommer
Beyond the Sum of Parts: Voting by Grouping Dependent Image Fragments

Multi-scale, category-level object detection in cluttered scenes is one of the long standing challenges of computer vision. The two leading approaches to this problem are sliding windows and voting methods, which are based on the Hough transform. Sliding windows scan over possible locations and scales and evaluate a binary classifier on each candidate window. The computational burden of this procedure is still daunting although various techniques have been proposed to alleviate the complexity issue. Rather than using a single, global descriptor for objects, Hough voting avoids the complexity issues by letting local parts vote for parametrized object hypotheses. Object shape is captured implicitly by modeling the relative location of various object parts w.r.t. the object center.

Despite the current popularity of the method, Hough voting has two significant weaknesses: i) (semi-)local parts are independently casting their votes for the object hypothesis and ii) intrinsically global object characteristics like shape are assumed to be a mere sum of local parts. This assumption is against the fundamental conviction of Gestalt theory that the whole object is different from the sum of its parts. And indeed, popular semi-local features have a large spatial support which results in overlapping sampling and mutually dependent descriptors. It is however possible to actually utilize these dependences by integrating shape-based contour grouping into the voting procedure. Therefore, mutually dependent parts are grouped while solving the correspondence problem jointly for all the dependent parts in a group.

I will review the main challenges that the most popular object detection algorithms are facing today and I will discuss how a shape-based approach can overcome some critical limitations of these methods.

Sinisa Todorovic
Recognizing Human Activities by their Space-Time Shapes

Recognition of human activities in video is one of the most challenging problems in computer vision with important applications, such as surveillance and human-machine interaction. Significant progress has been made in addressing this problem for short-term, repetitive, and punctual actions (e.g., walking, answering the phone) by modeling appearance and motion properties of the associated space-time points. However, there is a growing demand for richer video interpretations, e.g., in terms of reasoning about spatiotemporal constraints of simple activities comprising more complex ones, interpreting human activities that may not be observed due to occlusion, and predicting an actor's intent. These problems are challenging because of a large semantic gap between the required high-level reasoning, and the common point-based video representation. To overcome this semantic gap it seems that a richer low-level video representation is needed, lifting from the points to mid-level features, which would facilitate diverse computations of video understanding.

To this end, we observe that human bodies and objects that people usually interact with are spatially cohesive and characterized by locally smooth motion trajectories. Therefore, in the space-time video volume, they occupy space-time subvolumes, referred to as tubes. Since the human body consists of parts, each characterized by potentially distinct motion patterns over different time intervals in the video, a human activity can represented by a hierarchy of space-time tubes, referred to as activity shape.

In our recent work, we formalize activity shape as a spatiotemporal graph, and demonstrate its advantages for video interpretation. Specifically, access to video tubes can be provided by a number of existing approaches to multiscale spatiotemporal video segmentation. The resulting tubes can be conveniently organized in a spatiotemporal graph, where nodes correspond to the tubes, and edges capture their hierarchical, temporal, and spatial relationships. Given a set of training spatiotemporal graphs, we learn their archetype, i.e., model graph, and pdf's associated with model nodes and edges. The graph model adaptively learns from video data relevant video tubes, and their space-time relations for activity recognition. This advances much prior work that typically hand-picks the activity primitives, their total number, and temporal relations (e.g., allow only followed-by relations), and then only estimates their relative significance for activity recognition. Since activity shapes are closer to the symbol level, our work demonstrates that they could facilitate activity recognition, with significant reduction in the number of training examples.

John K. Tsotsos
Attending to Shape

There is more to the detection of 2D shape than the concatenation of piecewise linear segments. Similarly, there is more to attention than the selection of points and regions of interest or biasing computations to task knowledge. Do humans attend to shape? And if so, what are the requirements for a computational system to enable it to do the same? What would the benefits be? This presentation will begin by overviewing the Selective Tuning model of visual attention and a new shape detection framework. A brief summary of human behavioral and non-human primate neurophysiology relevant to these will follow.
A proposal for how the two may be linked computationally with answers to the above questions will close out the talk.

Rudiger von der Heydt
Contour grouping in the visual cortex

The visual system processes shape for many different tasks, to find, to recognize, to grasp, to compare etc. In every case it retrieves the relevant information from the millions of signals that stream in through the optic nerves. This amazing performance indicates powerful mechanisms for organizing the incoming information. The gestalt psychologists first pointed out that vision tends to organize elemental visual stimuli such as points and lines into larger perceptual units, or figures, according to certain rules. Much of this organization occurs independently of what the subject knows or thinks about the visual stimulus. This lecture will review recent neurophysiological studies on figure-ground organization and contour grouping in the visual cortex.

Song-Chun Zhu
Object Representation: Structure vs. Appearance and 3D vs. 2D - an Information Theoretical Perspective

What is the optimal representation for an object category that can be learned from a set of examples? How could such representation be coded by the neural system? In this talk, I will make an attempt to address these problems from an information theoretical perspective. First, I present an information scaling phenomenon that reveals a continuous transition between structure (shape) and appearance (texture) due to zooming, density, and stochasticity of motion. This leads to a concept called imperceptibility - the fundamental limit of what can be perceived from images. It will be used as a criterion for explaining the transitions between our models and perceptions. Secondly, I will present how we can pursue representations - hybrid templates that integrate structures and appearance as well as 3D and 2D descriptions. This gives a quantitative measure for various texton and texture elements as information gain in the learning process. Thirdly, I will show compositional models that account for variations of configurations. If time permit, I will also discuss the PAC-learning rate, i.e. how many examples are necessary for learning the representation to a given precision.

The talk is based on joint papers with Yingnian Wu, Zhangzhang Si and Wenze Hu.