Midwest Vision Workshop, December 16-17, 2014

Location: TTI-Chicago, 6045 S Kenwood Ave, Chicago

Schedule outline

- December 16:

11am-12:30pm talks
12:30-1:45pm lunch
1:45-3pm posters
3-4:15pm talks
4:15-5pm break
5-6:15pm talks
7pm dinner

- December 17:

8:30-9am breakfast
9-10:15am talks
10:15-11:30am posters
11:30am-12:20pm talks
12:20-12:30pm closing
12:30pm lunch

Notes:

- talks are 25 minutes, including questions
- all posters can be up throughout the workshop (both days)

Detailed schedule

● December 16

11am: **Opening remarks**

11:15am-12:30pm Talks: 3D inference
session chair: Qixing Huang (TTIC)

- Ayan Chakrabarti, TTIC: Low-level Vision by Consensus in a Spatial Hierarchy of Regions
  We introduce a generic computational framework for the estimation of physical scene value maps using local models that are expected to hold piecewise across the scene (like piecewise planar depth for stereo). Inference in this framework is carried out by a network of nodes, with each node reasoning about the validity of the local model in one of a dense, overlapping set of regions that redundantly cover the image plane, and all collaborating to produce a globally consistent scene map.

- Derek Hoiem, UIUC: Reconstructing complete 3D shape from one RGB-D image

Our goal is to recover a complete 3D model from a depth image of an object. Existing approaches rely on user interaction or apply to a limited class of objects, such as chairs. We aim to fully automatically reconstruct a 3D model from any category. We take an exemplar-based approach: retrieve similar objects in a database of 3D models using view-based matching and to transfer the symmetries and surfaces from retrieved models.

- Joseph Roth, MSU : Unconstrained 3D Face Reconstruction
  The input to our algorithm is an “unconstrained” collection of face images captured under a diverse variation of poses, expressions and illuminations. The output of our algorithm is a true 3D face surface model represented as a watertight triangulated surface with albedo.

12:30-1:45pm Lunch (catered at TTIC)

1:45-3:00pm Posters

3:00-4:15pm Talks: scene parsing
session chair: Derek Hoiem (UIUC)

- Mohammadreza Mostajabi, TTIC: Feeforward Semantic Segmentation with Zoom-out Features
  We introduce a purely feed-forward architecture for semantic segmentation. We map small image elements (superpixels) to rich feature representations, extracted from a sequence of nested regions obtained by "zooming out" from the superpixel all the way to scene-level resolution. This approach exploits statistical structure in the image and in the label space without setting up explicit structured prediction mechanisms, and thus avoids complex and expensive inference. Our architecture achieves new state of the art performance in semantic segmentation, obtaining 64.4% average accuracy on the PASCAL VOC 2012 test set.

- Satoshi Ikehata, Washington University in St. Louis: Structured indoor reconstruction
  We propose a novel indoor scene reconstruction algorithm. The approach analyzes a large indoor scene at a macro scale, and produces a structured scene representation.

- Michael Maire, TTIC: Reconstructive Sparse Code Transfer for Contour Detection and Semantic Labeling
  We frame the task of predicting a semantic labeling as a sparse reconstruction procedure that applies a target-specific learned transfer classifier to a generic deep sparse code representation of an image. Our classifier utilizes this deep representation in a novel manner: rather than acting on nodes in the deepest layer, it attaches to nodes along a slice through multiple layers of the network in order to make predictions about local patches.

4:15-5:00pm break, posters

5-6:15pm Talks: Learning
session chair: Jia Deng (U. of Michigan)

- Jason Corso, Univ. of Michigan: Learning Compositional Sparse Models of Bimodal Percepts
  Various perceptual domains have underlying compositional semantics that are rarely captured in current models. We suspect this is because directly learning the compositional structure has evaded these models. Yet, the compositional structure of a given domain can be grounded in a separate domain thereby simplifying its learning. To that end, we propose a new approach to modeling bimodal percepts that explicitly

relates distinct projections across each modality and then jointly learns a bimodal sparse and compositional representation. In a tabletop robotic manipulation environment, we jointly consider vision and speech, grounding the bidirectional generative and compositional model in language syntax, and demonstrate the potential of this new direction working with real data.

- Jason Rock, UIUC: Learning to regress images (with applications in superresolution, coloring, etc.)
  In image regression, we seek to regress images against images. For example, we might regress clean images against noisy images. Image regression problems are not usually attacked in a regression framework, because solutions must balance local accuracy with long-scale spatial constraints. We describe a method that learns to manage that balance by producing an optimization problem, whose solution is the solution to a particular image regression problem. Our method is learned from supervised training data, using functional gradient descent in a LEARCH-like framework. We provide experimental results on three standard problems: image denoising, intrinsic image estimation, and colorization.

- Saurabh Singh, UIUC: Learning to Find Landmarks
  We propose a general method to find landmarks in images of objects using both appearance and spatial context. We apply this method without changes to two important problems: parsing human body layouts, and finding landmarks in images of birds. Our method takes a group of landmarks, and uses contextual and appearance information to add another landmark to that group. The choice of landmark to be added is opportunistic and depends on the image; so, for example, in one image a head-shoulder group might be expanded to a head-shoulder-hip group but in a different image to a head-shoulder-elbow group.

7pm Dinner in Hyde Park (details TBD)

December 17

8:30-9am breakfast (catered at TTIC)

9-10:15am Talks: video and motion
session chair: Jason Corso (Univ. of Michigan)

- Sven Bambach, Indiana Univ.: This hand is my hand: Hand disambiguation in egocentric video
  Egocentric cameras are becoming more popular, creating large volumes of video in which the biases and framing of traditional photography are replaced with those of natural viewing tendencies. This paradigm enables new applications, including novel studies of social interaction and human development. Recent work has focused on identifying the camera wearer's hands as a first step towards more complex analysis. In this paper, we study how to disambiguate and track not only the observer's hands but also those of social partners. We present a probabilistic framework for modeling paired interactions that incorporates the spatial, temporal, and appearance constraints inherent in egocentric video. We test our approach on a dataset of over 30 minutes of video from several pairs of subjects.

- Xiaoming Liu, MSU: Efficient Motion-Saliency Detection and Its Application to Video Analysis
  This paper first presents an efficient motion-saliency detection (EMSD) algorithm that discovers saliency in a video by examining a set of random neighborhoods of pixels. We apply EMSD to the sports genre categorization and demonstrate superior efficiency and accuracy over the dense trajectories method.

- Matthew Walter, TTIC: Learning Articulated Motion from Visual Demonstration

This talk will describe a technique that learns kinematic models of a priori unknown articulated objects from depth imagery.

10:15-11:30am Posters

11:30am-12:20pm Talks:

- Jia Deng, Univ. of Michigan: Mining Semantic Affordances for Visual Object Categories
  We study the problem of mining the knowledge of semantic affordance: given an object, determining whether an action can be performed on it. Specifically, we connect verb nodes and noun nodes in WordNet, or equivalently fill an affordance matrix encoding the plausibility of each action-object pair.

- Yasutaka Furukawa, Washington University in St. Louis: Uncanny Valley for 3D Reconstruction
  Accurate 3D reconstruction is usually the key to high quality visualization applications. However, very often, improving reconstruction accuracy degrades the quality of visualization. This issue is little known to researchers, yet very important in practice.

12:20-12:30pm Closing remarks

12:30-2pm Lunch (catered at TTIC)

2pm departure

List of posters

1. Sven Bambach, Indiana Univ.: "This hand is my hand: Hand disambiguation in egocentric video"
2. Mohammed Korayem, Indiana Univ.: "PlaceAvoider: Steering First-Person Cameras away from Sensitive Spaces"
3. Jingya Wang, Indiana Univ: "Observing the natural world with Flickr"
4. Qixing Huang, TTIC: Creating Consistent Scene Graphs Using a Probabilistic Grammar
5. Greg Shakhnarovich, TTIC: Discriminative Metric Learning by Neighborhood Gerrymandering
6. Harry Yang, TTIC: Evaluation of vision algorithms by sensor prediction
7. Xi Yin, MSU: Jointly Multi-Leaf Segmentation, Alignment, and Tracking from Fluorescence Plant Videos
8. Amin Jourabloo, MSU: Attribute-preserved Face De-identification
9. Joseph Roth, MSU : On Continuous User Authentication via Typing Behavior
10. Joseph Roth, MSU: On the Exploration of Joint Attribute Learning for Person Re-identification

11. David Johnson, U. of Michigan: Semi-Supervised Nonlinear Distance Metric Learning Via Forests of Max-Margin Cluster Hierarchies

12. Chenliang Xu, U. of Michigan: Action Understanding with Multiple Classes of Actors

13. Yu-Wei Chao, U. of Michigan: Mining Semantic Affordances for Visual Object Categories

14. Joseph DeGol, UIUC: Recognizing materials in natural scenes

15. Liwei Wang, UIUC: Training very deep networks

16. Juan Caicedo, UIUC: Object localization

17. Micah Hodosh, UIUC: Natural Image Annotation and Search with Natural Language

18. Aditya Deshpande, UIUC: Mutli-stage SfM - Revisiting Incremental Structure from Motion

19. Taehwan Kim, TTIC: Reconstruction of missing fingerspelling motion capture data with a matrix completion approach

20. Mohit Bansal, TTIC: What are you talking about? Text-to-Image Coreference

21. Qixing Huang, TTIC: Estimating Image Depth using Shape Collections