Human Action Recognition and Shape Segmentation-Recognition

Jianbo Shi

Human Action Recognition. Human action recognition has broad range of applications such as video search, sports analysis, human robotics interactions, and health care. Our work is organized in two directions: 1) detailed pixel-level 'motion and pose', focusing on close interactions among people; 2) action recognition focusing on goal oriented motion, simplified as 'action = motion + intention'.

In "Detecting Unusual Activity in Video" (cvpr2004), we demonstrated that using large amount of un-labeled video data and a robust graph co-clustering approach, one can uncover visual patterns of un-usual and usual actions. This was an exciting discovery, as it suggested that big-data can solve this hard vision problem without explicitly defining action categories, and without detailed analysis of human motion. Through more experiments, it was clear that such big-data approach has an 'autistic' limitation: it memorizes many details, but understands little intricate relationships of human motion and causality among them. It has little ability to make long-range prediction of future actions. My recent work on human action recognition is aimed to resolve this 'autistic' limitation.

"Action = motion + intention".

Intention as And-Or graph of Actor-Actions: Storyline model. In "Understanding Videos, Constructing Plots: Learning a Visually Grounded Storyline Model from Annotated Video" (cvpr2009), we studied causality among actions. Our Storyline model can be regarded as a stochastic spatio-temporal grammar, whose language (individual storylines) represents potential plausible explanations of new videos in a domain. The basic insight is that all the variations of an event share a goal directed sequence of actions (akin to how movies of the same genre has similar flow of story subplots). The variations, such as one falls down as he runs towards a car, are due to different effects of each action. Our method requires only short video segments accompanied by a text description of the actors and actions present in the video. The system requires no detail annotations of the actor and actions in the video. We model the storyline grammar as a probabilistic AND-OR graph. The Storyline inference

outputs: 1. A set of full sentences describing all the actions of humans in the video. Its elements are nouns and verbs such as actors, actions, and objects words. They are connected through spatial-temporal relationship words such as before, after, during, next-to); 2. Visual appearance and casual relationship of these words. We applied storyline model on 39 videos clips of 2007 World Series. For the detection task of baseball events, our method nearly doubles the precision of the baseline methods 80% vs 40%, and has a much higher recall 85% vs. 10% for a state-of-art IBM language translation model).

Intention as Actor-Action-Goal Topological Planning. In "Multi-hypothesis Motion Planning for Visual Object Tracking" (iccv2011), we developed a long-term topological motion planning model to reason how people move in a crowded city. Persistent partial occlusions are the main challenge. We focused on constructing plausible goal-oriented plans using topological motion planning. We encode moving right and left of an obstacle, his movement "intention", using topological winding angles and winding numbers, and construct disjoint plans in different homotopy classes. For each person, our planner maintains multiple hypotheses for future paths as they move in the environment, creating a virtual simulation of intended pedestrian motion. When a person is visible, we track him, and use his trajectory to narrow down the set of plausible goals/planned paths. When a person becomes occluded, we create multiple hypotheses that predict his re-appearance based on the plausible set of goals/planned paths provided by the planner.

"Motion and Pose".

In "Pose from Flow and Flow from Pose" (cvpr2013), we developed an algorithm that mediates the information between body parts recognition and multi-frame motion grouping to improve both human pose detection and tracking. Our approach produces pixel-level body part segmentation, detection, and motion flow. We focused on the challenging cases when both motion flow and pose estimation fail at the same time, which occur often when body motion is large or pose is in an uncommon state. We ask the question, "Can bad motion information be useful?" Our observation is that body motion, though not accurate, is often sufficient to segment the body parts from background. Once body parts are segmented from background, by matching these segments to known shape exemplars, one can improve pose estimation robust to uncommon posture. We also ask "Can imprecise pose estimation be useful?" We recognize that pose estimation, though not precise, can provide better pre-conditions to motion estimation by adapting the model based motion models, making it robust to background clutter. We show significant improvement (2x accuracy) on tracking fast moving lower arm, which achieves the additional goal of pixel-level labeling for the body parts.

The technical foundation of this mediation approach is based on "Graph Steering", which is developed in "Two-granularity tracking: Mediating trajectory and detection graphs for tracking under occlusions" (eccv 2012). The main challenge is to reason multiple representations of actions jointly, while dealing with uncertain and often conflicting informations. We utilized two representations for our study, first on coarser-level body detection with discriminative trained people classifiers, and second on fine-level dense optical flow computed using image brightness gradient. We constructed a two level graph to encode people detection and motion tracking as a joint grouping problem. The challenge is not only how to combine informations but also figure out which representations to trust, and when. Our 'graph steering' computation actively modifies the graph weights on the joint graph locally, so to reach a global consensus of detection and tracking, using joint segmentation on this graph as a consensus and confidence measure.

Shape Segmentation-Recognition. Shape is an expressive abstraction of visual patterns in natural images. In computer vision, many applications can benefit from accurate shape recognition, including robotics, image search, video analysis and medical image understanding. Shape is a critical cue for recognition, as it is sufficiently invariant to represent commonalities of different instances of a particular object category, while preserving enough detail about objects in order to differentiate them from each other or the background. It also varies systematically with 3D viewpoint, enabling estimation of the object pose. While there are many different approaches to use object shape for recognition, there are two difficulties faced by nearly all approaches: object deformation and background clutter.

We have focused on the three sub-tasks of object recognition: 1) Detection: indicating the presence or absence of an object at a particular location in the image; 2) Alignment: determining the pose of an object by corresponding it to a shape model, and 3) Segmentation: determining the boundaries of the object, necessary for manipulating it and interacting with it.

Through our experiments, we identified two quantities (two ratios) affect our ability to see shape in complex images:

1. **Deformation Ratio**: for a category of object shape, it is the ratio between a) the deformation between the *rigidly detectable* parts and b) the *size* of these parts. The uncertainties of the deformation can be un-isotropic. For 'shape' objects, such uncertainties of deformation are often very large in an unknown 1D space. Most of the success we have seen on object detection has been on objects with small "deformation ratio", and with uncertainties of deformation well constrained in a 2D domain isotropically.

2. **Clutter Ratio**: for an image, it is the ratio between a) size of the object, and b) size of the segmentable region on object. We can allow sampling in the segmentation space to produce multiple segmentations, so long it is not too large. For semantic scene recognition, we have seen success when large object shapes can be segmented using bottom-up cues.

The challenge is working with an object category with large deformation ratio (related to detectable parts), and an image with large clutter ratio (large uncertainties in segmentation due to clutter).

Our earlier work has focused on shape as 'Deformable Graphical Model and Co-Segmentation-Recognition', where pair-wise attributes are used for reasoning geometric relationship and are encoded using an attributed graph. We realized an approach based on a fixed size object parts can not handle the two challenges mentioned above.

Shape Packing: a Jigsaw Model.

We developed holistic shape matching using bottom-up image structures such as image contours and segments, for object shape detection, alignment and segmentation. Holistic shape matching utilizes global information about object shape, rather than local image features which often contain too little information to match reliably to the object model.

We start with salient bottom-up image segments or contours, and actively reason about image fragments which occur in unpredictable ways (depending on context). Because of the unpredictable image fragmentation, there is no one-to-one correspondence between image structures and object parts. We developed an approach for *many-to-many* matching formulated as a 'Packing' problem. Given two sets of contours (or segments), the goal was to find a subset of contours (or segments) that had similar *holistic shape*. Shape similarity was measured by comparing shape contexts computed over the selected subsets of contours, and a computationally efficient approximation to this combinatorial problem was formulated as a linear program 'Packing' problem. The many-to-many matching was used to detect object parts in the image, which was then combined via a voting scheme to provide object detection scores. The approach was evaluated on the ETHZ Shape Classes dataset, achieved the top performance with 91.1% average precision (AP) and near perfect detection rate at 0.004 false positives per image (FPPI).

The key innovation is to construct a holistic shape descriptor that is an *algebraic* function of the latent selection variable of foreground and background contours. As such, we can use computational tools for the combinatorial Packing problem as a robust computational solution. This is a sequence of work published in 'Many-to-one contour matching for describing and discriminating object shape' (cvpr2010) and 'Discriminative Image Warping with Attribute Flow' (cvpr2011).