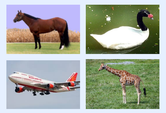
Discovering Object Classes from Activities

Abhilash Srikantha[1,2] and Juergen Gall[1]

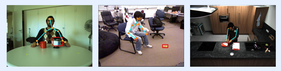
[1]Computer Vision Group, University of Bonn
[2]Perceiving Systems Department, Max Planck Institute for Intelligent Systems

1.Quick Summary

- Object models require a vast amount of training data to perform well
- Recent shift of attention to utilize weakly annotated data in videos
- Fundamental assumption of present day methods:
  - Motion and/or appearance of the object of interest is dominant
  - Object of interest forms the main theme of the video



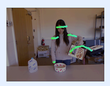
- **Problem: Work on small and medium sized objects**
  - Video data for objects like mugs, plates etc. is scarce
  - Labelled human activity data available in plenty
  - Previous assumptions do not hold: dominant human



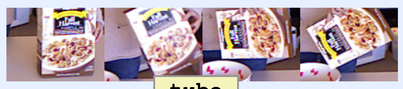
- **Input to the system**
  - Set of videos of similar activities
  - Automatically extracted Human Pose
- **Output from the system**
  - Object tubes common to all videos
  - One tube per video
- **Datasets for Experiments**
  - ETHZ (RGBD, TOI, Model Based Pose est.)
  - CAD-120 (RGBD, Kinect, OpenNI tracker)
  - MPII-Cooking (RGB, Pictorial structures)
- **Conclusions**
  - Appearance insufficient for small objects
  - Big gains from encoding Functionality
  - Present day pose estimation is good enough

2.Tubes Generation

$$p(T_v) = \sum_S \sum_\tau p(T_v|\tau, S)p(\tau)p(S)$$



random seeds

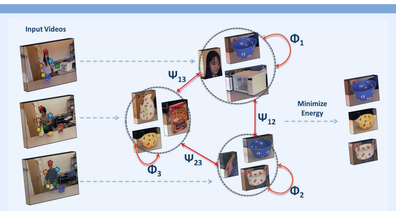


tube

- Tubes generated by randomly selected superpixel and tracking algorithms

3.Model

Input is a set of action videos with human pose. Instances of the co-mmon objects are discovered by defining similarity in apperance and functionality as:



$$E(L) = \sum_v \Phi(l_v) + \sum_{v,w} \Psi(l_v, l_w)$$

Unary: App (saliency) | Body avoidance | Pose-object-relation | Size

Binary: Shape | Functionality

APP | FUN | SIZ

$$\Phi(l_v) = \lambda_1 \Phi^{app}(l_v) + \lambda_2 \Phi^{pose}(l_v) + \lambda_3 \Phi^{body}(l_v) + \lambda_4 \Phi^{size}(l_v)$$

$$\Psi(l_v, l_w) = \lambda_5 \Psi^{shape}(l_v, l_w) + \lambda_6 \Psi^{func}(l_v, l_w)$$

4.Unary Terms

- Appearance saliency: chi^2 RGB(D) distance between inside a tube and around it

$$\Phi^{app}(l_v) = \frac{1}{K}\sum_{k=1}^{K}\left(1 - \frac{1}{2}\sum_i \frac{(I_{k,i} - S_{k,i})^2}{I_{k,i} + S_{k,i}}\right)$$

- Pose-object relation: median distance between closest joint and center of the tube

$$\Phi^{Pose}(l_v) = \frac{1}{K}\sum_{k=\alpha\cdot K}^{(1-\alpha)\cdot K} \|c_{D(k)} - j_{D(k)}\|$$

saliency

object-pose relation

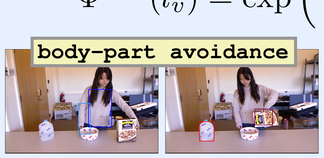
- Body avoidance: maximum response of the body-appearance/skin model

$$\Phi^{body}(l_v) = \max\{\bar{p}_{skin}(I), \bar{p}_{upper}(I), \bar{p}_{lower}(I)\}$$

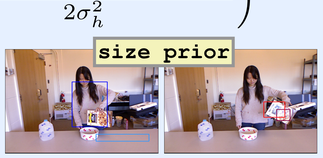
- Size: Variation of object size based on the size of the hand

$$\Phi^{size}(l_v) = \exp\left(\frac{(w_{l_v} - 2w_h)^2 + (h_{l_v} - 2h_h)^2}{2\sigma_h^2}\right)$$

body-part avoidance



size prior



5.Binary Terms

- Functionality: (Normalized) Head-Object distance after DTW alignment

$$\Psi^{func}(l_v, l_w) = \text{median}_k\left\{|d_{\omega_v(k)} - d_{\omega_w(k)}|\right\}$$

functionally dissimilar

functionally similar

- Shape: Median PHoG distance between frames after DTW alignment

$$\Psi^{shape}(l_v, l_w) = \text{median}_k\left\{\frac{1}{2}\sum_i \frac{\left(P_{\omega_v(k),i} - P_{\omega_w(k),i}\right)^2}{P_{\omega_v(k),i} + P_{\omega_w(k),i}}\right\}$$

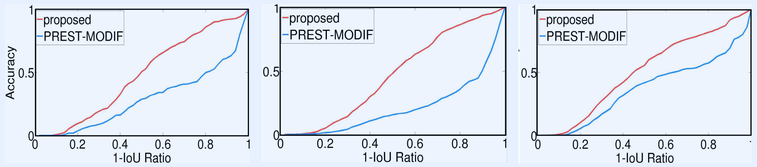
dissimilar shape

similar shape

6.Inference

- Model parameters set using Validation Dataset (one class per dataset)
- Use loopy belief propogation (TRW-S) algorithm for inference

7.Results



IOU distribution: Cumulative frame IOU distribution for MPII, ETHZ and CAD-120

Comparison with state-of-the-art: Prior art full model, prior art using proposed tubes, full proposed model

	prest-exact	prest-modif	proposed
ETHZ	0.063	0.249	0.447
CAD	0.039	0.246	0.410
MPII	0.023	0.221	0.342

	proposed	APP	APP+SIZ	FUN	APP+FUN	FUN+SIZ
ETHZ-Action	0.447	0.192	0.305	0.292	0.312	0.390
CAD-120	0.410	0.168	0.191	0.147	0.202	0.350
MPII-Cooking	0.342	0.079	0.149	0.229	0.235	0.288

Evaluating potential groups: Average class IOUs for various combinations

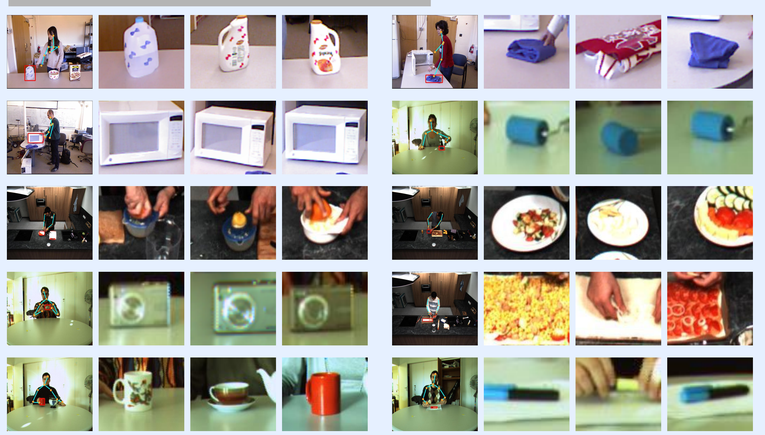
	Φ^{app}	Φ^{pose}	Φ^{body}	Φ^{size}	Ψ^{shape}	Ψ^{func}
ETHZ-Action	0.35	1.88	-25.49	-13.50	-4.62	-8.86
CAD-120	-48.66	-15.73	-18.89	-20.80	-40.15	-9.19
MPII-Cooking	-15.85	0.06	-31.09	-10.70	0.058	-60.95

Evaluating individual potentials: (%) change in average class-IoU when discarded

	ETHZ	CAD	MPII
GTruth	60.6	29.4	47.8
Inferred	53.2	24.4	35.3

Comparing object models: Average precision (%) of object detectors from groundtruth and inferred tubes

8.Inferred Tubes



asrikantha@informatik.iai.uni-bonn.de
gall@informatik.iai.uni-bonn.de