2013 AMALTHEA REU Program

Unified Framework for Interaction Recognition

Jacob Chen[1], Haidar Khan[2], Ivan Bogun[3], and Eraldo Ribeiro[4]

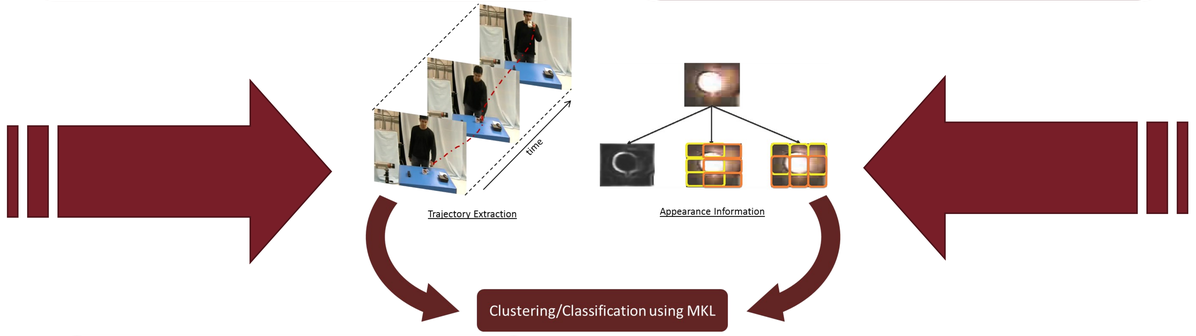
1. jchen114@terpmail.umd.edu, 2. hkhan47@hawkmail.newpaltz.edu, 3. ibogun2010@my.fit.edu, 4. eribeiro@cs.fit.edu

Problem Statement

Interaction Recognition is an important and open problem in Computer Vision. The problem involves using video data to identify interactions between actors (humans) and objects. The complexity of the problem arises from the large amount of variation within a class of interaction. Interactions in the same class vary by the objects used in the action, the motion(s) involved, and even the actors.

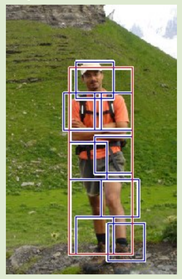
General Approach

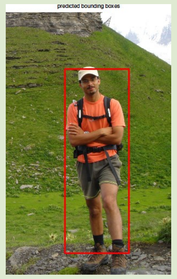
We propose a framework to identify human-object interaction by combining several types of information about the action. We gather motion information in the form of trajectories using tracking by detection. We obtain appearance-based information about the objects using PLSA. Finally, we combine and classify interactions using MKL Methods



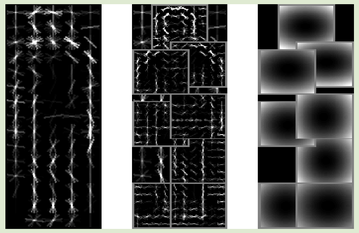
Motion Based Information

Object Detection





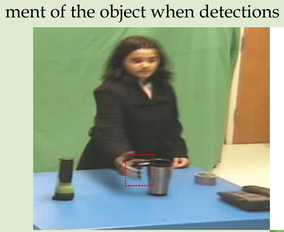
In order to extract spatio-temporal information from videos, we must be able to detect key objects. To accomplish this, we trained deformable part based models based on [1]. We trained models to detect hands and the full body.

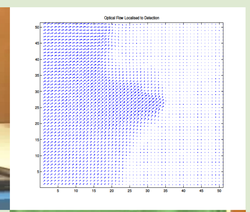


Human detector model. Left: Root Filter. Center: Part filters. Right: Deformation costs for each part filter.

Trajectory Extraction

To actually construct trajectories, we used a modification of tracking by detection which was augmented by optical flow. Our algorithm relied on localized optical flow to estimate the movement of the object when detections were not available.





Optical flow calculates the movement of pixels in successive frames and when localized to a detection, it can be used to estimate the next position of a trajectory.

Comparing Trajectories

- We normalized the generated hand trajectories w. r. t. the body trajectory
- To compare trajectories, we used the Dynamic Time Alignment Kernel from [2]

$$K_1(x,y) = \underset{\pi \epsilon A(x,y)}{\operatorname{argmax}} \frac{1}{|\pi|} \sum_{i=1}^{|\pi|} \exp\left(\frac{\| x_{\pi_1^{(i)}} - y_{\pi_2^{(i)}} \|^2}{-\sigma^2} \right)$$

Multiple Kernel Learning

Multiple Kernel Learning allows us to combine complementary motion and appearance information

We can combine our kernels using:

$$K(\cdot,\cdot) = \sum_{i=1}^{n} \gamma_j K_j(\cdot,\cdot)$$

In order for this combination to be positive semi-definite, $\{\gamma_j\}_{j=1}^n$ should represent a convex combination. We chose a simple heuristic to choose these weights as a normalized similarity to the ideal kernel:

$$\gamma_j = \frac{A(K_j, yy^T)}{\sum_{i=1}^{n} A(K_i, yy^T)} \quad \forall j = 1,\ldots,n$$

Experiments/Results

- Leave-One Out Cross-Validation

	Annotated % Success	Automatically Generated % Success
Trajectories	75.0	51.85
Velocities	90.7	46.29
Full Framework	96.2	-

Action Classification using SVM

We implemented a One vs. All Multiclass Support Vector Machine to classify the actions.

The dual form of the optimization problem for soft-margin SVM is given as:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j k(x_i, x_j)$$

$$\text{s.t. } 0 \le \alpha_i \le C, \quad i = 1,\ldots,m$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0$$

Due to the limited nature of our dataset, we used Leave-One Out Cross-Validation (LOOCV) and trained a classifier for each type of action.

Future Work

- Improved tracking and detection
- Interaction localization
- Complete automation

References

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, Object detection with discriminatively trained part based models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1627[1645, 2010.

[2] H. S. K.-i. Noma, Dynamic time-alignment kernel in support vector machine," vol. 2, p. 921, 2002.

[3] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, Discovering objects and their location in images," vol. 1, pp. 370[377, 2005.

[4] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis," Machine learning, vol. 42, no. 1-2, pp. 177[196, 2001.

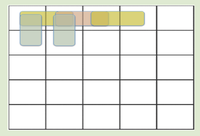
Appearance Based Information

Objective

- Object recognition in videos
- Learn similarity between videos with respect to objects

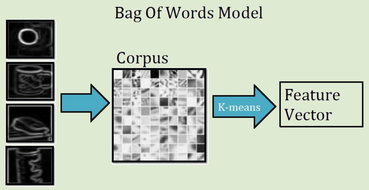
Feature Extraction

- Edges
- Overlaying Patches

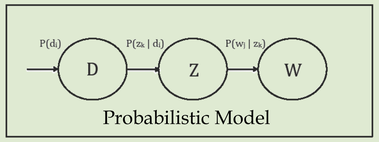


Approach

- Bag Of Words model [3]



- Probabilistic Latent Semantic Analysis [4]



- Learn probabilistic distribution by implementing Expectation-Maximization algorithm
- Expectation Formula:

$$P(z_k \mid d_i, w_j) = \frac{P(w_j \mid z_k) P(z_k \mid d_i)}{\sum_{i=1}^{K} P(w_j \mid z_i) P(z_i \mid d_i)}$$

- Maximization Formula:

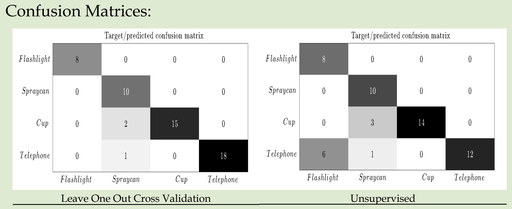
$$P(w_j \mid z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j) P(z_k \mid d_i, w_j) + \gamma_j}{\sum_{j=1}^{M} \gamma_j + \sum_{m=1}^{M} \sum_{i=1}^{N} n(d_i, w_m) P(z_k \mid d_i, w_m)}$$

- Build a kernel which describes the similarity between videos with respect to objects:

$$K(r_i, c_j) = \sum_{k=1}^{M} P(z_k \mid d_i) P(z_k \mid d_j)$$

Experimental Results

- We tested many combinations and determined the best one by consistency and accuracy level
- Best combination was the propagation of information from Doublets to Edges
- Confusion Matrices:



This material is based upon work/research supported in part by the National Science Foundation under Grant No. 1263011.