Improved Hierarchical Deep Temporal Model for Group Activity Recognition

Mostafa S. Ibrahim, Alaa Eldin Abdelaal, Minlong Lu and Hanhan Wu

Simon Fraser University

Problem

- Group activity recognition attracts attention of vision community, specially human activities
- Scene Collective activity: Most frequent activity by individuals in a scene
- One of recent work is based on deep learning by Ibrahim et al. We are improving their work.

Our Contributions

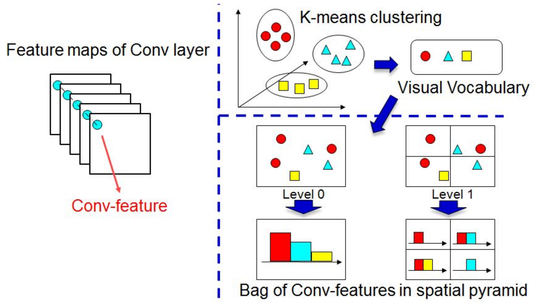
- A **Spatio-temporal graph** representation for the scene to encode spatial informations and their temporal dynamics
- A **Spatial Pyramid BOVW model** based on CNN convolutional feature maps (SP-BOC).
- A **systematic approach** to little fixing low accuracy classes in NN setups.
- Explore using **CNN** features over **unstructured feature** representation from the first hierarchy.
- Using **Random Forests** to classify 2nd stage temporal data instead of the soft max layer
- Investigate potential areas application of 2-stage model.

Spatio-Temporal Graph Repres.

- The motivation is to decrease the confusion between walking and crossing.
- Build a graph for every frame with nodes representing the position of each person at a given time (hence the name “Spatio-temporal”).
- Design a new feature vector (temporal data from the 1st stage + spatial data from the graph) and feed it to a NN classifier. Push up the walking results by 1%.

Spatial Pyramid Bag of Conv-features (SP-BOC)

- Conv-feature: concatenation of the values in the same position of the convolutional layer feature maps
- SP-BOC representation: partition the image into increasingly fine sub-regions and compute the histograms of Conv-features found inside each sub-region
- Classification using SVM with histogram intersection



Second Stage Enhancements

- Instead of feeding unstructured vector from stage to the 2nd network. We **reshaped** input as 170x170 and feed to network similar to alex one. However, performance was ~ 15-20% accuracy across different network structures. Lesson of the day, **CNN can’t work over unstructured data.**
- Another trial is to **fix low accuracy** classes in final models is to **train for little** steps over low class, then to the overall classes in **repetitive manner**. Push up from 0-2%
- We extract the 2nd layer LSTM data and feed them to **random forests,** aiming at decreasing the error rate with the maximization of the strength and minimization of the correlation. Push up 0.5 %.
- We investigated further areas for the application of 2 stage model, it can be promising in **Medical Sensor areas** which deal with time-based medical features analysis.

Experiments

Model Name	Accuracy %
Ibrahim et al - Main Model	81.5
PushUp Trick	81.7
RandomForestsStage2	**82**
Spatio-Temporal Graph	77.6
CNN for layer 2	17

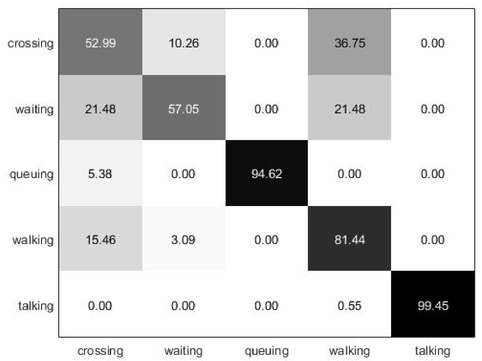


Table 1: Confusion matrix for the spatio-temporal graph representation.

Training Data Selection	Cross Validation Average Accuracy - Decision Tree using Recursive Partitioning		Average Accuracy - Random Forest
Non-Temporal Selection	76%		79%
Temporal Selection	Train on all the frames	78%	**82%**
	Train on each 6 frames	74%	

Table 2: Random Forest Results

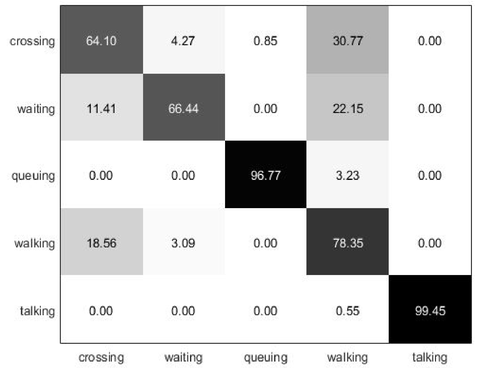


Table 3: Confusion matrix for the Push Up NN Trick. We got little improvements.

The SP-BOC method is evaluated on the **volleyball dataset.** The Conv-feature is extracted from the Alexnet Conv5 layer in the Baseline 1 - image classification.

Model Name	Accuracy %
Ibrahim et al - Baseline 1	46.7
Ibrahim et al - Main Model	51.1
SP-BOC	**55.88**

Table 4: SP-BOC result on volleyball dataset