Machine Learning

Group Activity Recognition

Mostafa S. Ibrahim

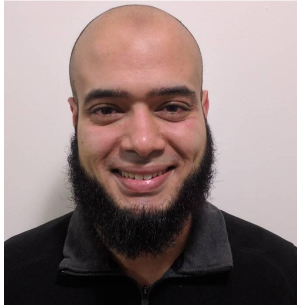
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher

PhD from Simon Fraser University - Canada

Bachelor / MSc from Cairo University - Egypt

Ex-(Software Engineer / ICPC World Finalist)

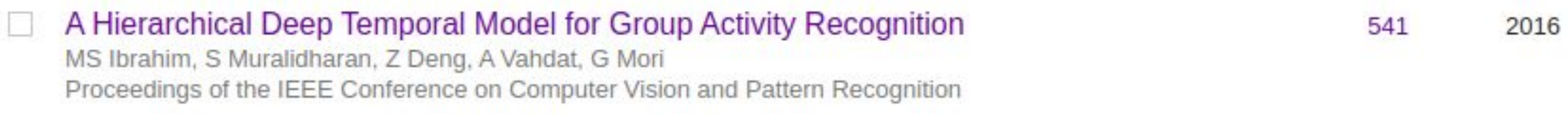


© 2024 All rights reserved.

Please do not reproduce or redistribute this work without permission from the author

A Hierarchical Deep Temporal Model for Group Activity Recognition

- In 2016, I published a CVPR paper with an important dataset
- The model represented a deep learning treatment to this problem
- The dataset is a reference for the problem
- Today, March 2024, the paper is cited 500+ times



□ A Hierarchical Deep Temporal Model for Group Activity Recognition 541 2016
MS Ibrahim, S Muralidharan, Z Deng, A Vahdat, G Mori
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

Problem and Description

- Papers: CVPR and Extension
- Official C++/Caffe implementation (and Dataset)
- Video in Arabic and its slides

Guidelines

- Understand the video
- Read the **extension** paper (1-3 times)
  - There is an improvement from CVPR paper after using group-style
- Properly understand the dataset
  - Be able to load and visualize it
  - Print its statistics
- Build simple baseline first

Dataset

- In practice, your first step is to fully understand your data
  - Quality and Quantity
  - Print all possible relevant statistics
  - Visualize a random sample
  - Write down observations on the properties of the data and bias in it
- Download the dataset: 60G dataset
  - Or start with a sample of 2 videos each with 2 clips
- Understand the dataset properly (internal notes / githhub)
- This dataset has 2 levels of annotation
  - 9 person actions
  - 8 scene classes

Some Support

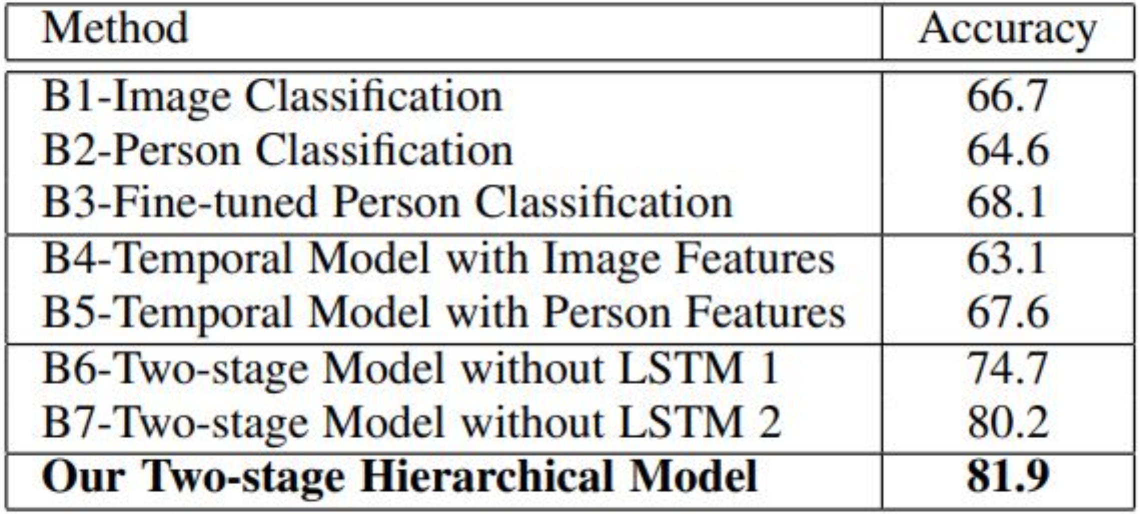
- You are supposed to code the project fully
- I built 2 useful scripts for you
    - Script that can visualize the whole dataset
        - It can also save a pkl file for the annotation
    - A script that creates resnet50 model and use it to extract features
        - Either image level or box level
- Please debug both scripts line by line and fully understand
    - This will be strong base
    - Feel free to skip them and build them by yourself

Ablation Study

- A method to assess the impact of various **components** of a system on its overall performance by experimenting by removing
- Example: Assume your system consists of 4 enhancement features
    - Let’s name them: A, B, C, D
        - Example A and B are 2 extra losses. C is a 2nd LSTM layer. D is complex backbone
        - Then you do experimentations such as
            - ABC, ACD, ABD. Each one will tell you the effect of a single component removal
    - A classifier: Removing certain layers of the neural network, disabling data augmentation techniques, or using different feature extraction methods
    - In self-driving car: removing sensors to see the effect

Volleyball Ablation Study

- Each experiment teaches you something.
  - For example, naive image classification doesn't help
  - For example, temporal information is boosting the results
    - LSTM #1 has high performance impact on the model
  - Using one representation per team reduced the confusion and enhanced the results



Method	Accuracy
B1-Image Classification	66.7
B2-Person Classification	64.6
B3-Fine-tuned Person Classification	68.1
B4-Temporal Model with Image Features	63.1
B5-Temporal Model with Person Features	67.6
B6-Two-stage Model without LSTM 1	74.7
B7-Two-stage Model without LSTM 2	80.2
Our Two-stage Hierarchical Model	**81.9**

Baseline B1-tuned

- Don’t try anything that doesn’t finetune (well-proved idea)
- In CVPR paper I used alexnet. You better network (e.g. **resnet50**)
- For each clip, use the middle image only
  - Fee free to use 5 before and 4 after also
- Fine-tune an image classifier over 8 classes
- Compute the results. This is your first model

Baseline B3

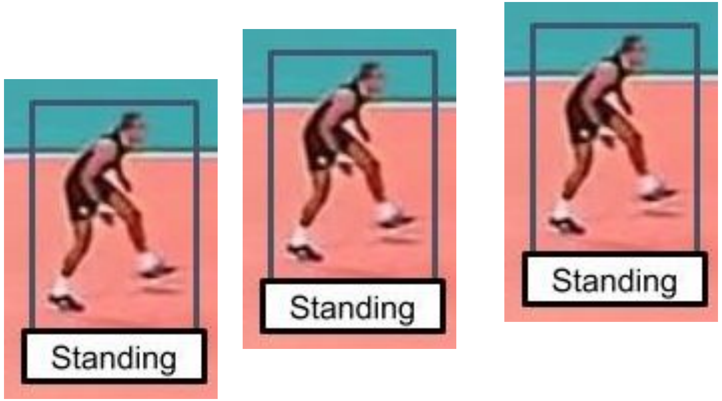
- A) Train: Fine-tune an image classifier over 9 actions
  - Input is a cropped person
- B) Inference: For an image
  - Get all the persons crops
  - Feature extraction for each crop, e.g. 2048 features
  - Max pool all the features = now this is an image representation
- C) Train: Do NN training on these features over 8 classes

Baseline B4

- Implementation #1
    - Use the classifier of B1-tuned to extract representation per clip
        - Use 9 frames per image
        - Now you have sequence for each clip of 9 steps
    - Now train an LSTM on these sequences
    - Start with this implementation.
- Implementation #2
    - Extend the classifier network directly with an LSTM layer then classification
    - This way no need to do explicit features extraction

Baseline B5

- Temporal on crops (LSTM on player level)
- Similar 2 implementation paths to build representation per person
- You can represent each clip with the last hidden state
- Max pool all players representations (9 per image)
- Then do the NN network exactly like in B3 (on images)
    - Features classifier - no temporal info



Baseline B6

- Same B3 steps A and B
    - For B, you will extract representations for each clip of 9 frames
- For C
    - Do LSTM on sequences from step B
- This is a model where LSTM is applied on the image level only

Baseline B7

- Full model V1
- A) train LSTM on crops level (LSTM on a player)
- B) extract clips: sequence of 9 steps per player
- C) for each frame, properly max pool its players
- C) train LSTM 2 on the frame level

Baseline B8

- Same as B7
- The scene representation is not pool of all players
- X = Pool team 1 6 players
- Y = Pool team 2 6 players
- Let scene representation concatenation of X and Y

More

- Implementation Challenge (a side from memory issue)
    - Can you implement the 2 stage model as a single network
        - Input is 12 cropped users (make sure each person is a track of a single person)
        - LSTM on person level and LSTM on scene level
            - You need 3 modules; LSTM 1 - pool properly - LSTM 2
        - Loss on person action and loss on scene classification
- You can learn GNN and extend the network
    - My 2018 paper: Hierarchical relational networks for group activity recognition and retrieval
- There are a lot of papers on this problem
    - Learn and apply
- Study the literature of this problem. It will help you see how ideas advances
- Problem with similar techniques: motion prediction for cars in self-driving

"Acquire knowledge and impart it to the people."

"Seek knowledge from the Cradle to the Grave."