**Pose Recognition from Single to Sequential Video Frames**

|  |  |
| --- | --- |
| Deren Lei  Undergraduate  Computer Science  derenlei@ucsb.edu | Zichen Sun  Undergraduate  Computer Science  zichen\_sun@umail.ucsb.edu |

**Abstract**

*Google’s recently released dataset AVA II focusing on activity recognition densely annotates 80 atomic visual actions in 430 15-minute video clips. With its realistic scene and action complexity, AVA II[3] exposes the intrinsic difficulty of action recognition. The current public focus is on designing time sequence models like Fisher vector encoding to analyze extracted features. Our research project will focus on doing experiments directly on static video frames regardless of the time sequence.*

**1. Introduction**

The activity labels in AVA II we are interested in are the poses without interactions with subjects. Specifically, actions with labels of stand, sit, walk, run, get up, etc. We will analyze these actions beginning with single video frame in AVA II[3] and JHMDB[2] dataset. Then labeling acions from sequential video frames via CNN-LSTM model on JHMDB[2] dataset. We will compare the results of single video frame between AVA II[3] and JHMDB[2] dataset, and the results of single and sequential video frames in JHMDB[2] dataset.

**2. Experiments**

**AVA II single video frame pose recognition**

a) Extract video frames via YouTube ID stored in dataset.

b) Corp images by person\_box given in dataset.

c) Build a basic CNN structure for action recognition and test it on coped images of humans.

d) Use the same structure on complete video frames.

c) Compare the results.

**JHMDB II**[2] **single video frame pose recognition**

a) Extract video frames via YouTube ID stored in dataset.

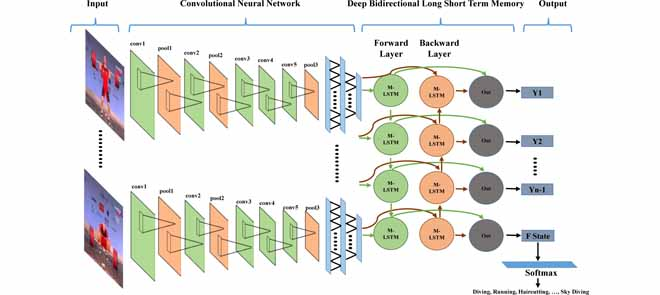
b) Use the same structure implemented in the first experiment to test on this dataset.

c) Compare the result with the first experiment.

**JHMDB II**[2] **sequential video frame pose recognition**

a) Extract video frames one per second for each video.

b) Build a CNN-LSTM model[1][4] and test it on the sequential video frames.



c) Compare the result with the second experiment.

**3. Future research**

Test our CNN-LSTM model on AVA II dataset when the dataset with person label is released.

**4. References**

[1] ZHENG, Zhenxing. “Multi-Level Recurrent Residual Networks for Action Recognition.” *Multi-Level Recurrent Residual Networks for Action Recognition*.

[2] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A Large Video Database for Human Motion Recognition. ICCV, 2011.

[3] Gu, Chunhui, et al. “AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions.” *AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions*.

[4] Ullah, Amin. “Action Recognition in Video Sequences Using Deep Bi-Directional LSTM With CNN Features.” *Action Recognition in Video Sequences Using Deep Bi-Directional LSTM With CNN Features - IEEE Journals & Magazine*,

ieeexplore.ieee.org/document/8121994