Assignment for EE6222 AY 2023/2024

Human Action Recognition in the Dark: a Simple Exploration with Late Fusion and Image Enhancement

Background:

Video data is one of the most common forms of data widely used in all aspects of our daily life. With the rapid growth of video data (500 hours of videos uploaded to YouTube daily alone), automatic video analysis has become a crucial task in dealing with these vast number of videos. Among various video analysis tasks, **human action recognition (HAR)** is one of the cornerstones, which aims to recognize (classify) a human action automatically. The emergence of various large-scale video datasets, along with the continuous development of deep neural networks have vastly promoted the development of HAR, with increasing application in diverse fields, e.g., security surveillance, autonomous driving, and smart home.

Despite the rapid progress made by current HAR research, most research aims to improve the performance on existing HAR datasets constrained by several factors, one of which concerns the fact that videos in existing datasets are shot under a non-challenging environment, with adequate illumination and contrast. This leads to the observable fragility of the proposed methods, which are not capable to generalize well to adverse environments, including dark environments with low illumination. Take security surveillance as an example: automated HAR models could play a vital role in anomaly detection. However, anomaly actions are more common at nighttime and in dark environments, yet current HAR models are obscured by darkness, and are unable to recognize actions effectively. It is therefore highly desirable to explore methods that are **robust** and could cope with dark environments.

Project Description:

The project aims to guide you to explore HAR in videos shot in the dark leveraging the late fusion technique. This project also aims to encourage further exploration on how image enhancement effects HAR in videos shot in the dark. You will be given a set of training data, which includes 25 random videos for each of the 6 action classes, and a set of testing/validation data. Your task is to classify the testing data into one of the 6 action classes by training a classifier and evaluate the performance of the classifier. You will be guided through a detailed guideline as listed in the following section, with steps that you are advised to follow. You are required to submit a full report of strictly no less than 4 pages in ICLR format which should demonstrate your process of exploration and answer the questions in detail that are listed in the guideline. You are also advised to observe the additional directions listed at the end of this document.

Detailed Project Guideline:

Section 1: Step 1 – Frame Sampling. To perform HAR, we begin with sampling video frames such that the feature extraction is more feasible. The two common sampling strategies are uniform sampling and random sampling. Pick any video as an example and demonstrate the difference/similarity between the different sampling strategies. Discuss which strategy should be considered first. (3 points)

Section 2: Step 2 – Feature Extraction. With the sampled frames, we then obtain the features of each video. A straightforward approach is to leverage the late fusion strategy where we first obtain the feature of each sampled frame and fuse them just before the classifier by a pooling (average) operation performed across all sampled frames. To obtain the feature of each sampled frame, it is recommended to leverage pre-trained models that are trained on large datasets and possess good generalizability. Prior to obtaining the features of each frame, you should **normalize** the pixel values such that the pixel values are of zero-mean and unit standard deviation. The reference mean and standard deviation value of the dark video frames is $\text{mean}\,[0.07, 0.07, 0.07], \text{standard deviation}\,[0.1, 0.09, 0.08]$. Describe in brief the pre-trained model leveraged and why the pre-trained model is selected. What is the dimension of the feature obtained. Remember to save the video features in order for subsequent training. (3 points)

Section 3: Step 3 – Classifier Training and Evaluation. Select a feasible classifier (e.g., SVM, Bayes, MAP, etc.) and train the classifier with the obtained feature and labels provided. Discuss the pros and cons of the type of classifier selected. Subsequently, evaluate the trained classifier. You should repeat steps 1 and 2 for the validation videos to obtain their features and obtain their class predictions with the trained classifier. Compare the predictions with the ground truth label. What is the performance of the trained classifier? (3 points)

Section 4: Step 4 – Effects of Leveraging Image Enhancements. For us humans, applying image enhancements could significantly improve our capability in recognizing actions performed in dark videos as image enhancements could produce much clearer video frames. Does our HAR model follow such intuition? Apply any image enhancement of your choice and explore how it effects the performance of the trained classifier. Note that the reference mean, and standard deviation value of a normal video frame is $\text{mean}\,[0.485, 0.456, 0.406], \text{standard deviation}\,[0.229, 0.224, 0.225]$. Discuss how the chosen image enhancement effects the performance of the trained classifier in detail. Provide sampled output frames resulting from the image enhancement. (6 points)

Section 5 (Optional): Step 5 – Improving the HAR Model to Enable End-to-end Training. The aforementioned method is intuitive but is not end-to-end, which limits its applicability in real-world scenarios. Currently, most HAR models are designed end-to-end, without the need to explicitly store the video features. In this step you are to design or implement an HAR model that is end-to-end and evaluate your HAR model. Describe your HAR model in detail, including the structure along with the training and evaluation procedures. Compare your HAR model performance against the prior trained classifiers and discuss the pros and cons of your HAR model. (Additional 10 points)

Appendix: Code – Include your code as screenshots in the Appendix. (-5 points IF NOT INCLUDED)

Additional Directions and Notes:

1. The report must be submitted in PDF format, following the provided ICLR Conference template (in LaTeX). You may use the ICLR Conference Template on Overleaf (for free).
2. The report must be done individually. You may discuss with your peers, but NO plagiarism is allowed. The University, School, and the teaching team take plagiarism very seriously. An originality report may be required from iThenticate (available on NTULearn) when necessary. A zero mark will be given to anyone found plagiarizing and a formal report will be handed to the School for further investigation.
3. There will be 6 action classes (jump, run, sit, stand, turn, walk) in the provided data. The training, and validation/testing set will all be provided along with files that state the ID of the video, the class ID of the video and the video file name.
4. You may use any method deemed necessary. There are no limitations to what framework/library should be used.
5. It is highly recommended to program your method with **Python** or **MATLAB**. However, other programming languages are also acceptable. You are required to submit your code as an Appendix in the final report.
6. The **deadline** for this Assignment is Monday on Week 14 (Revision Week 1), 20 November 2023, at 1659 GMT+8.
7. Name your report strictly as: "**YourFullName_YourMatriculationNo..pdf**" (if your full name as in your matriculation card is Xu Yuecong and your matriculation number is U2000000K, then your submission pdf file is named as "Xu Yuecong_U2000000K.pdf"). Submit your report to NTULearn.
8. Penalty marks will be applied for late submission or if you do not follow the above submission guideline.