Project Proposal: Emotion Recognition in Videos Using Deep Neural Networks

Koushik Roy (A20478667)
Deep Learning
Instructor: Martin Hagan

March 28, 2025

1. Problem Selection and Motivation

The problem selected for this project is **emotion recognition in videos using deep learning models**. Accurately identifying human emotions is a critical challenge in computer vision and human-computer interaction. Traditional emotion recognition systems typically rely on static images, which lack the temporal dynamics present in real-world emotional expression. In contrast, videos offer richer context and subtle transitions in facial expressions that are crucial for detecting nuanced emotions.

This problem is chosen for its relevance in fields such as mental health assessment, virtual assistants, autonomous systems, and surveillance. By developing a model that can analyze emotions from facial expressions in videos, we aim to create a system that mimics the human ability to perceive emotions over time.

2. Dataset

We will use the **RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)** dataset. It contains over 1,400 video clips from 24 professional actors displaying eight emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. Each clip is less than 5 seconds long with a resolution of 1280x720. We will downsample the videos by extracting 6, 10, or 16 evenly spaced frames per clip and resize them to 224x224 pixels to reduce computation.

This dataset is large enough to train deep neural networks, and the labeled emotional expressions make it highly suitable for supervised learning tasks.

3. Network Architecture

We plan to implement and compare the following deep learning architectures:

- **Early Fusion 2D CNN:** Frames are concatenated along the channel axis and passed through a 2D CNN.
- **Late Fusion 2D CNN:** Each frame is processed independently through a CNN, and feature maps are concatenated before classification.
- **3D CNN:** A 3D convolutional network that jointly learns spatiotemporal features from the frame sequences.
- **RNN-CNN:** A hybrid model where CNN extracts spatial features and an LSTM layer models temporal dependencies.

While some of these models are standard, they will be customized in terms of frame input size, layer configuration, and fusion strategy to suit the dataset and problem.

4. Framework and Tools

The models will be implemented using the **PyTorch** framework. PyTorch provides dynamic computational graphs, modular design, GPU acceleration, and a rich set of tools for vision-related tasks, making it an ideal choice for this project.

Supporting tools:

- `torchvision` for data augmentation and preprocessing
- `Matplotlib/Seaborn` for visualization and analysis
- AWS EC2 or a local GPU-enabled setup for training

5. Reference Materials

We will reference the following literature and tools:

- Fan et al. (2016), "Video-based emotion recognition using CNN-RNN and C3D hybrid networks"
- Simonyan and Zisserman (2014), "Two-Stream Convolutional Networks for Action Recognition"
- RAVDESS dataset documentation [1]
- PyTorch official documentation and tutorials
- Course materials on CNNs, RNNs, and 3D convolution

6. Performance Evaluation

Model performance will be judged using the following metrics:

- **Accuracy** on a held-out test set
- **F1 Score** (macro-averaged), to account for class imbalance and better reflect performance across all emotion categories
- **Confusion Matrix**, to analyze which classes are frequently misclassified
- **Per-class accuracy**, to evaluate recognition performance for specific emotions
- **Training time and model complexity**, as secondary considerations

We will also compare models under different data conditions, such as varying frame counts and with/without data augmentation.

7. Project Timeline (4 Weeks)

Week	Milestone
Week 1	Dataset preprocessing, frame extraction, resizing, and train/val/test split
Week 2	Implement Early Fusion and Late Fusion models; begin training and testing
Week 3	Implement and train 3D CNN and RNN-CNN models; compare all model performances
Week 4	Perform evaluation, analyze results (F1 score, confusion matrix), finalize report and presentati

8. Conclusion

This project explores the application of deep learning architectures to emotion recognition in videos. By comparing multiple neural network approaches (2D CNNs, 3D CNNs, and RNN hybrids), we aim to gain insights into the effectiveness of temporal modeling in video-based emotion classification. Performance will be evaluated using accuracy, F1 score, and confusion matrices. In future work, we hope to incorporate audio data for multimodal emotion recognition and explore transformer-based architectures for longer sequence modeling.

References

[1] Livingstone, Steven R., and Frank A. Russo. *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS).* PLoS ONE, 13(5), 2018.