A Review of Multi-Modal Emotion Recognition Approaches for Affective Computing

- Anish Pimpley

Presented by **Anish Pimpley**
for CS 527@ UMASS

Project Type:
Literature Review

Summary

Topic	To review the development and current cutting edge of multi-modal models for emotion recognition.
Goals	● To trace major developments in the field chronologically ● To identify dominating theories between SOTA models ● To discuss and compare fusion techniques used for combining models ● To highlight challenges, criticisms and possible avenues for improvement
Reason for choosing this project	● I was a great learning opportunity to freely explore the multimodal recognition landscape. ● The explosion of data, computational power and deep learning has disproportionately benefited recognition more than other sub-domains of affective computing. ● The interdisciplinary nature of the domain, makes reviews and consolidated reports contextualizing the field, more important than in other domains. ● Existing studies were either too broad, dated or only looked at uni-modal analyses.
Conclusions	● Discussed methods for feature extraction ,fusion and prediction for the problem. It's beginnings, changes in popular methods and presently dominant architectures. ● The lack of standardization across benchmarking methods and datasets needs to be addressed by the field. ● Clear trend towards using deep learning methods in both feature extractors and fusion methods. ● Suggested subfields (3D ML, VQA) from where this field could borrow ideas

Take home message

- Deep Learning methods are rising to prominence, but classical methods are still competitive
  - SOTA models are moving towards end-to-end networks enabling joint training.
  - Classical models such as SVMs, hand engineered features, OpenSmile, HMMs can achieve similar results after some ensembling and fine tuning using domain knowledge.
- The results from SOTA models like C3D-DBN, Tzikaris et al. and CRMKL show significant improvements over classical models. The steady improvement of Deep Learning model performance is encouraging.
- The lack of standardization across datasets and benchmarking methods makes it difficult to compare results
- There is no single dominant neural network architecture. However cariants Spatio-temporal CNNs and LSTMs form the core of most architectures.
- SOTA models treat individual networks as black boxes. No much work is being done on modifying network structures themselves.
- Audio, visual and textual modalities get a lot more attention than other physiological, gesture or pose modalities.
- SOTA models still borrow heavily from classical models like C3D-DBNs(2017) from Kim et al.(2013) and CRMKL (2015) using kernel methods that were popular in the 2000s.
- Strides in 3D deep learning and VQA (both vision sub domains dealing with multiple modalities) could be utilized to make breakthroughs in affective computing.
- Multi-modal would greatly benefit from a unanimously agreed upon benchmarking standard and a common dataset to verify model generalization against.

Personal Experience

- The project was a lot of fun. It was certainly not easy as some of the more complex neural network techniques sent me down a rabbit hole of background readings.
- However, the classical methods were easy to understand. This is because many of them were either covered in class or I was more familiar with them from past experience with ML, NLP or Vision.
- The main takeaway was how despite narrowing down the search space to deep learning based multimodal models using A-V-T for multi modal emotion recognition, I still found myself surrounded by a huge amount of literature, that I could never reflect on too deeply in the review.
- The rapid takeover of the problem by deep learning was also interesting to witness. The end to end nature of deep learning makes it significantly easier to pose the problem as a black box, streamlining the models while still giving great results.
- It is also interesting to see how feature extractors such as FACS and OpenSmile still manage to hold their own. It is testament to the quality of these models.
- I would like to change my approach from being chronological (inspiration -> effect) to one based around datasets. Restricting the search space this way would have caused me to miss out on some major models. But, I would have been able to compare models in a more direct manner, since they would share a common benchmarking framework.

Thank You.