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

Project Report for UMass CS527: Affective Computing

Anish Pimpley
MS in CS at UMass Amherst
apimpley@cs.umass.edu

Abstract

Affective computing lies at the intersection of AI, NLP, vision, social and cognitive science disciplines, with each having distinct approaches to the field. Most of these disciplines have observed major technological transformations in the last decade, heralded by the exponential rise of data availability, computational power and the rise to prominence of machine learning. The ready availability multiple modalities in the form of synchronized video, audio, text and sensing information has disproportionately benefited efforts to solve the emotion recognition problem. The field has accordingly moved to develop techniques that leverage these modalities and their interactions. Most previous studies have exclusively reviewed models for individual modalities or been too broad to explicitly focus on this problem. In this review, we trace the trajectory of the multi-modal emotion recognition problem, from its origins to the distinct modern approaches that represent the current state-of-the-art. We particularly focus on the approaches used to fuse modalities and strides on the deep learning front that have been utilized to improve emotion recognition results. Lastly, we identify parts of the problem space that are still considered difficult and possible avenues for pushing the field forward. The review aims to help practitioners from diverse backgrounds to get a comprehensive overview of the best solutions posed by different approaches for this problem.

Introduction

Affective Computing as field covers the broad area of using intelligent computational systems for detecting, recognizing, interpreting and expressing human emotion. Emotion recognition is a sub-domain within affective computing. It is an active area of research within the vision, NLP, social science, cognitive science and AI research community.

Since the rise of social media, and video streaming websites like Youtube, a massive amount of data is available in the form of audio, video, and not text alone. For instance, a post about a person visiting a location is often accompanied by a high definition video of event with synchronized audio. In addition, the rise of smartwatches means that certain physiological data of the event may also be preserved. While video is innately a richer medium than text or images, it presents information along multiple modalities. The frames store information along the spatial and temporal domain. The video has a corresponding audio channel and there are established methods to extract the textual content of the same video. Classical methods have focused on unimodal models, which underutilize information from every other input channel. As human beings, we too rely on multi-modal signals [1]. Secondly, each person utilizes each mode to different degrees for expressing emotions [2]. This means that an audio-based model would struggle to recognize emotions in a person with an expressive face and bland tone, and vice versa for an imaging-based model. In that sense, unimodal models using different channels struggle on different problems, and no single channel contains enough information to accurately recognize all emotions. In addition to this problem, unimodal systems are plagued by the notorious problem of missing data.

It is well documented, that multimodal models outperform unimodal models. [8] Given the context of videos, this review shall focus on the fusion of audio, visual and textual modalities.

Fig.1 represents a generic multimodal emotion recognition architecture for the audio-visual inputs.

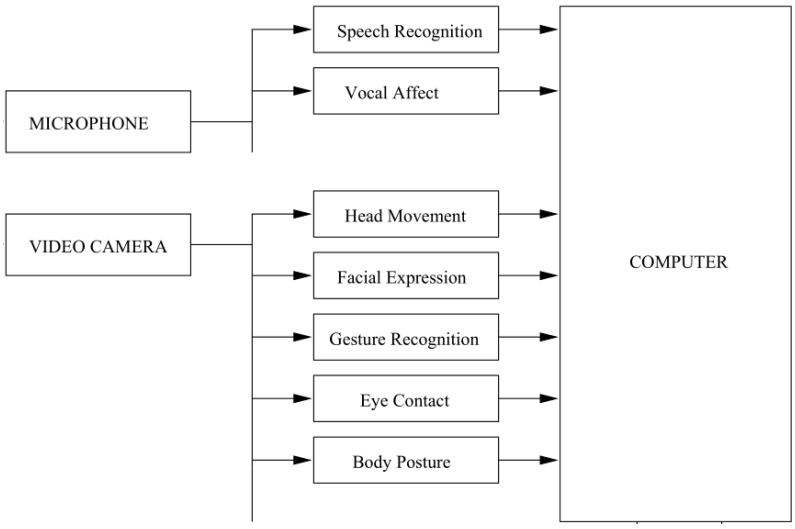
Related Work

Poria et al. [3] compared the strengths of multi modal systems over unimodal systems in a all-encompassing overview of affective computing. However, the large scope of the study restricted them for focusing on a specific problem or model family.
Sebe et al. [4] published a review addressing the exact problem as ours. However, the paper is now dated and does not address the drastic changes observed over the last decade. Recently, a modern review of emotion recognition systems using temporal features of video was published. However, it does not take audio, textual, or non-facial features into account.
The closest analogue to our study, is by D'Mello et al. [5], but its focus lies on comparing data collection, representational methods and benchmarking multimodal systems against unimodal systems.

Selection Process and Inclusion Criteria

Models were selected by searching through recently published and well cited papers at top conferences. Winners of recent multimodal emotion recognition competitions [6][7], best papers & honorable mentions at emotion recognition workshops were given priority as well. We required that models be benchmarked on known prominent datasets for easier and clearer comparison. We prioritized models that focused on leveraging big data and new developments in machine learning, over ensemble based or hand tuned models. Thus, deep learning based systems formed most of the selected models.

Fig.1 recognition system for a generic multimodal system taking audio-visual inputs. [5]



Multi Modal Emotion Recognition

Early models

The first multimodal emotion recognition frameworks can be traced back to the early 2000s. [8][9] [10]. These systems relied on the then standard facial and aural feature extractors to independently retrieve image and audio level features. These studies introduced 2 types of fusion methods for combining the extracted audio and video features.
Feature level models combined the extracted features first, and then computed scores for detecting emotions on the newly combined common feature set. Decision level models independently recognized the emotion, and then combined scores to provide a final prediction.

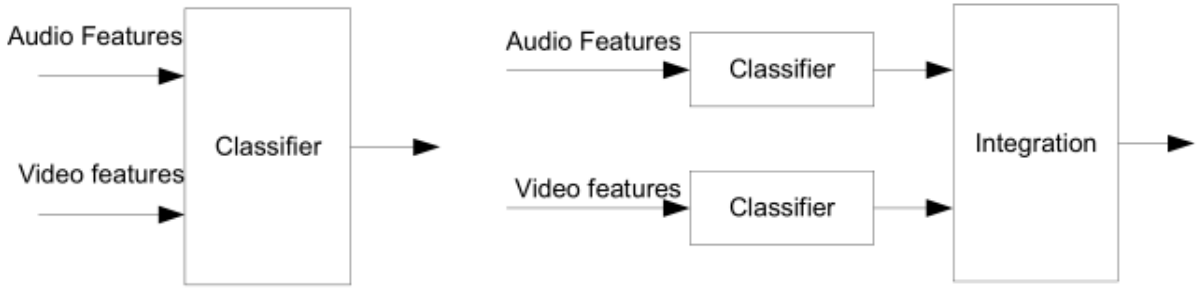


fig.2: fusion (left) and decision (right) based early multimodal systems

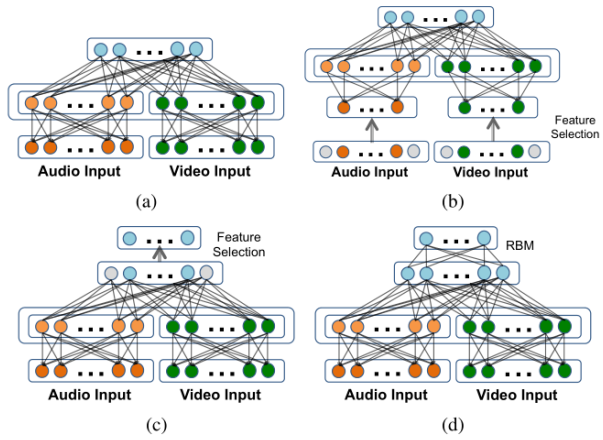
The fusion of both scores and features could be computed using common operations such as majority voting, maximum, average combining, product combining or weighted combining. Product and weighted combining delivered best overall results. However, the authors noticed that the benefits of multimodal systems different across different emotions.
Each of these models were designed in an era of low data availability. Thus, the feature extractors were not data driven, and hand tuning of feature and combination weights was a lot more common.
In following years, features extractors improved as methods like forward selection, information gain, PCA became widely used. Pitch, energy, mel-frequency filter banks and statistical functions of raw audio were the most commonly used auditory features. Models also performed better when features were normalized on a per-speaker basis. [11]

These systems modelled richer features. However, the nature of such models meant that they could not exploit non-linear relationships between each modality. These factors limited the capacity of early systems.

Deep learning based approaches

Deep Belief Networks

Deep Learning and Neural Network models rose to prominence in machine learning community following the introduction of [12] CNNs and [13] RNNs that cemented themselves as state of the art in domains with spatial and temporal information respectively. However, the first major deep learning model for multimodal emotion recognition came in the form of Deep Belief Networks (DBNs) [14] by Kim et al. in 2013.
DBNs can be viewed as a stack of RBMs where the feature space is represented as a distribution over hidden nodes, and the graphical structure denotes the dependence of each node on parent audio-visual inputs or features.



#	Model	Accuracy
(a)	DBN2	70.46
(b)	DBN2-FS	72.96
(c)	**DBN3**	**73.78**
(d)	FS-DBN2	72.77
Baseline 1	IG -SVM	73.38
Baseline 2	PCA - SVM	70.02

Fig.3 DBN variants (left); Table 1. Performance of DBNs on IEMOCAP data (right)

Models based on DBNs allowed for end-to-end feature learning, while maximizing the likelihood of the observed data. The stacking of RBMs allowed the model to capture nonlinear relationships between audiovisual features, while still preserving desired constraints using the graph structure of the model. The full RBM model (c) was able to outperform every model despite not using any hand-engineered features or unsupervised feature extraction methods.

Convolutional Deep Belief Networks

Ranganathan et al. [15] built on this by introducing a Convolutional Deep Belief Network, where RBMs are replaced by convolutional RBMs in the video pipeline of the network. They also extend the DBN to take audio, visual and physiological inputs by introducing the emoFBVP dataset, which has data for all 3 channels.

Feature level fusion using Convolutional MKL

Recently, deep learning model architectures have moved away from generic catch-all models like DBNs and instead specialize depending on their applications. Most significantly, CNNs and LSTMs have become the defacto networks for extracting visual and time series features respectively. Thus, we see a return to feature level fusion models, where audio, visual and textual features are computed independently of each other, and then fused before the decision block.

However, a key difference in the newer models, is that all feature extractors are deep learning architectures, and the model learn both features and the classifier end-to-end. Fig. 4 shows the current state of the art published by Poria et al. [16]

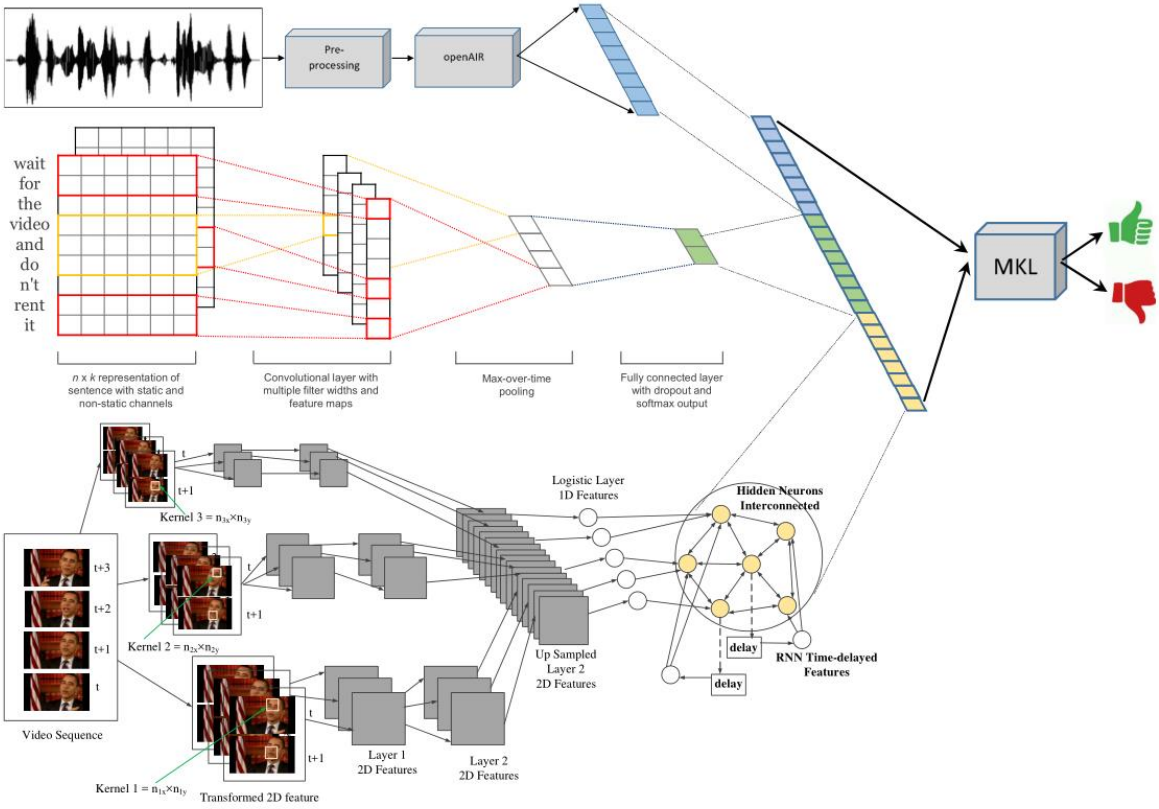


fig. 4 CRMKL model combining sentiment features in audio, video and text

The CRMKL consists of 4 parts. Namely, the visual, text, audio feature extractors and the MKL module. The visual feature extractor is a Convolutional RNN the computes hidden features for every 2 consecutive frames and considering them as one image. Interestingly, the textual

feature extractor uses a CNN to extract features from text. However, the text is presented as a concatenation of a word-2-vec word embedding and the 6 possible parts of speech tags. Contrary to the 1st two parts, the audio feature extractor does not use machine learning or neural network, instead using OpenSmile [17] which instead captures prosodic and statistical properties of the audio file. The extracted features are concatenated and passed to the multiple kernel module, which can be viewed as a generalization of the kernel property exploited by SVM. Instead of using a single kernel for all modalities as in an SVM, MKL enables each modality to have a unique kernel assigned to it, while still learning a non-linear decision boundary.

Model (A,V,T modalities)	Angry	Happy	Sad	Neutral
Ensemble of SVM trees [18]	78.1	69.2	67.1	63.00
DBN3 (audio + visual modality)	73.78 (averaged)			
CRMKL	79.2	72.22	75.63	80.35

Table 2: Performance against the then state of the art model and DBNs on IEMOCAP

Temporal fusion

The winner of the recent EMOTIW challenge [19] used a similar base network for their challenge as well. However, they replaced the MKL module with a weighted mean module. Also, they use temporal fusion (fig. 5) ie. overlapping features from adjacent frames and an extra VGG Face feature extractor to further improve their results.

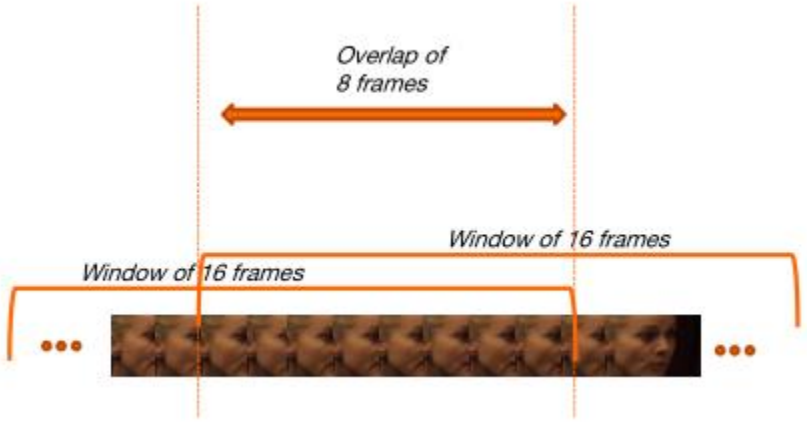


Figure 6: temporal fusion

Jointly trained End-to-end deep learning methods:

All systems discussed till this point have used some form of feature extractor in at least one modality to build their model. In contrast to these methods, jointly trained models take raw input for all modalities as input and train the feature extractors and the fusion weights jointly, all at once. Particularly, the auditory channel is also modelled with an ML method, instead of statistical property extractors like OpenSMILE.

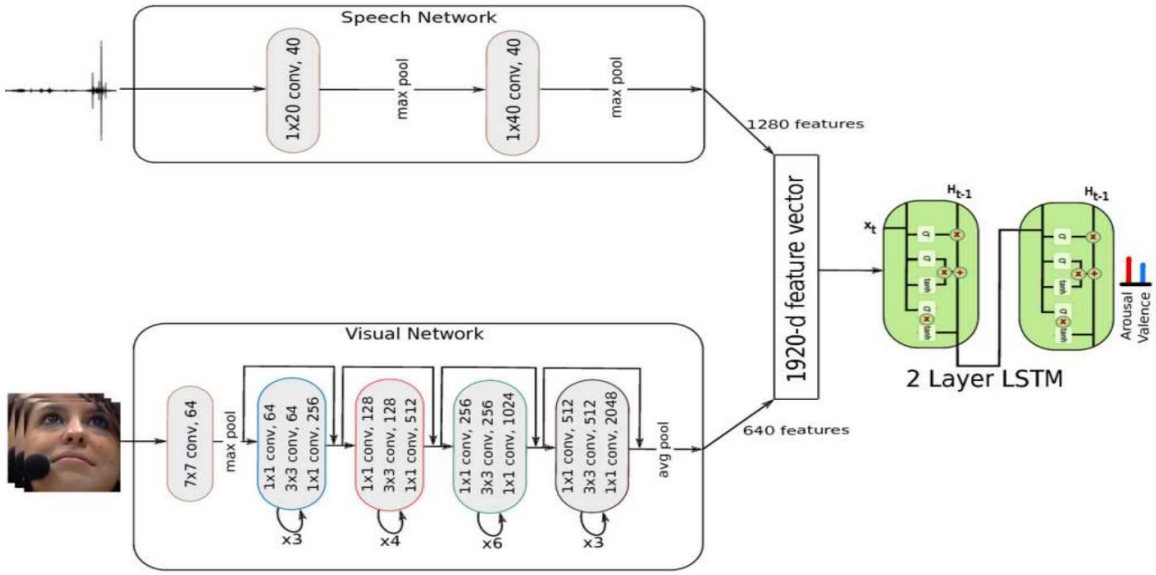


Figure 7: end to end network by Tzirakis et al.

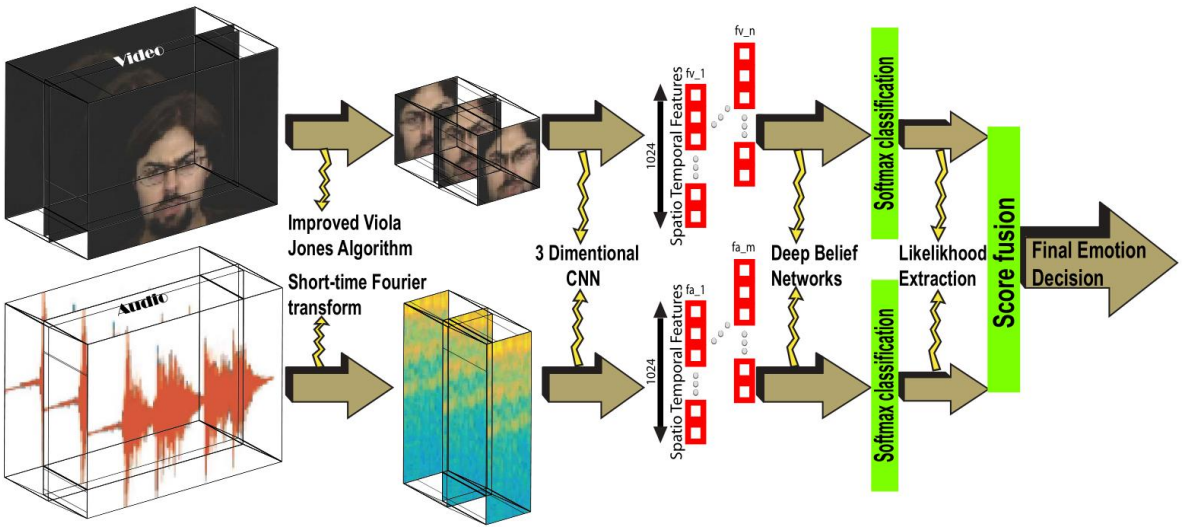


Figure 8: Deep spatio-temporal features for multimodal emotion recognition

Above are two such networks. Tzirakis et al. [20] use a 2 layer LSTM to fuse features extracted by a speech and visual network independently. The speech network uses segmented

waveforms as input, and then apply temporal convolutions and pooling across time to retrieve audio features. The visual network is derived from the 2016 ImageNet winning model ResNet. The key innovation by Tzirakis et al. was to use an LSTM to capture contextual information during fusion, as each of the common feature is fed to the network in an orderly fashion. Results are reported on the RECOLA dataset used for the AVEC challenge 2016. The competing models both use strongly hand tuned features and internal model structure. They do not use neural networks in any capacity. The results on RECOLA differ from previous results, in that the target predictions are values of the concordance correlation coefficient for valence and arousal. Thus, while results cannot be directly compared, we can put them in perspective against its competing models

Model	Arousal (correlation coef.)	Valence (correlation coef.)
OARVM-SR [21]	.**770**	.545
Han et al. [22]	.610	.463
Tzirakis et al.	.714	.**612**

Table 3: RECOLA dataset results (in terms of ρ c) for prediction of arousal and valence

The Deep spatio-temporal feature model by Nguyen et al. heavily leverage modelling guidelines from its predecessors. Raw data is preprocessed unlike Tzirakis et al. The face bounding boxes are extracted using Viola-Jones cascade classifier and the audio stream is transformed into its corresponsing DFT before passing to the network. The model uses 2 C3D networks [23] to extract visual information. The auhors suggest that the network far outperforms any 2D frame based CNN. The use of C3D on audio features is a first by the authors. The key innovation was to pass the extracted features to a DBN as seen in Kim et al. The authors introduce a new fusion method referred to as score level fusion. Instead of using feature-level or decision level fusion, the fusion step it done when computing the likelihood of the RBM instead. The results are inpressive, as they outscore the closest competing model on the ENTERFACE [24] test set by close to 10% accuracy points.

Method	**Anger**	**Disgust**	**Fear**	**Happiness**	**Sadness**	**Surprise**	**Avg.**
Feature and decision fusion approach [25]	-	-	-	-	-	-	39.00
Decision-level fusion [6]	-	-	-	-	-	-	56.27
Hybird fusion approach [21]	73.00	69	69.00	70.00	70.00	73.00	71.00
Feature level [24]	87.00	75.00	60.00	99.00	64.00	41.00	71.00
Score-level bimodal SVM [5]	-	-	-	-	-	-	87.40
Bayes sum rule (BSR) [28]	77.5	80.92	80.19	82.23	78.38	82.44	80.28
The proposed method (A-V C3D + DBN)	90.68	89.32	90.00	90.00	89.47	89.81	**89.39**

Discussion:

Observations:

We explored the landscape of multi modal emotion recoginition through a narrow lens, to trace its origins, chronological developments and methods that dominate the cutting edge of research in the domain.
We observed a clear direction in which the field is headed. Classical machine learning and linguistics models such as HMMs, semantic parsers, static feature extractiors, and SVMs still see significant use in research and hold their on benchmarks, newer developments are almost entirely being heralded in by neural networks. This is ofcourse no surprise, as the closest fields to this problem, Vision and NLP have almost entirely transitioned to deep learning based models.
There still appears to be no concensus in the research community, when it comes to feature fusion methods. Decision and feature level is still popular and provides competitive results. On the other hand, a lot of creative fusion techniques have been proposed in the form of RBMs, LSTMs and MKL modules.
Most models appear to either focus on video and text or video entirely. There is a concerning lack of models that leverage gesture, pose, physiological data in multimodal recognition. The lack of datasets to facilitate the same is identified as the root cause of this.
The lack of standardization appears to be plaguing the field. It is common for researchers to use lesser known metrics and underused datasets. This is an expected and understandable problem, as the interdisciplinary nature and fast moving pace make it very difficult for the community to agree upon rigourous standards unanimously. There also do not appear to be restrictions additional plugins that may be used to enhance a model. For instance the ML communites often avoid use of ensembles in addition to their model when comparing against other stock models. There is also a concerted effort to maintain common preprocessing and augmentation methods across all competing models. While this is not a major problem, it does make it difficult to compare across 2 wildly different approaches to the same problem.

Future possibilties:

Most of the explored models treat each neural network module as a block box. This restrcts the extent to which the researcher can manipulate its architecture to suit their problem. Another rising field in ML that deals with multiodal data is Visual Question Answering (VQA). It is common in VQA research to construct a network from scratch with a custom architecture. Concepts such as text based confitioning, neural attention dyanmic modules have led to state of the art results in VQA. Each of those models use invetive ways to represent interactions

between the visual and textual medium. I hypethsize, that adopting methods from the VQA research community may be of immense benefit to multimodal emotion recognition.
The addition of an ImageNet like equivalent for multi modal emotion recognition may benefit the field drastically. Unlike the vision problem, constructing such a dataset will a lot more difficult for emotion recognition, however its benefits cannot be overstated.
We expect the field to continue leveraging state of the art vision and NLP models to improve the feature extraction pipeline. There have been numerous strides in domain of feature extraction from videos in the past 2 years. The community may also look at this community for avenues of innovation.
As data from multiple modalties gets more common, it would be interesting to see models that account fo missing data, and are built around it. GMM and PGMs in general have obtained strong results in this domain.
Apart from the aforementioned criticisms, the field has a healthy amount of innovation and new deep learning techniques continue to improve outcomes. That being said, the challenges do imply that the key problems of the domain are far from solved.

References:

1. Shimojo, S. (2001). Sensory modalities are not separate modalities: plasticity and interactions. Current Opinion In Neurobiology, 11(4), 505-509. doi:10.1016/s0959-4388(00)00241-5
2. Morency, L., Mihalcea, R., & Doshi, P. (2011). Towards Multimodal Sentiment Analysis: Harvesting Opinions from The Web. International Conference On Multimodal Interfaces (ICMI 2011).
3. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, 98-125. doi:10.1016/j.inffus.2017.02.003
4. Sebe, N., Cohen, I., Gevers, T., & Huang, T. S. (2005, January). Multimodal approaches for emotion recognition: a survey. In Internet Imaging VI (Vol. 5670, pp. 56-68). International Society for Optics and Photonics.
5. D'mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. ACM Computing Surveys (CSUR), 47(3), 43.]
6. Dhall, A., Goecke, R., Joshi, J., Hoey, J., & Gedeon, T. (2016). EmotiW 2016: video and group-level emotion recognition challenges. Proceedings Of The 18Th ACM International Conference On Multimodal Interaction - ICMI 2016. doi:10.1145/2993148.2997638
7. Hugo Jair Escalante, Victor Ponce-Lopez, Jun Wan, Michael A. Riegler, Baiyu Chen, Albert Clapes, Serigio Escalera, Isabelle Guyon, Xavier Baro, Pal Halvorsen, Henning Muller, Martha Larson. "ChaLearn Joint Contest on Multimedia Challenges Beyond Visual Analysis: An overview", ICPR workshop 2016.
8. Chen, L.S., Huang, T. S., Miyasato T., and Nakatsu R. Multimodal human emotion / expression recognition, in Proc. of Int. Conf. on Automatic Face and Gesture Recognition, (Nara, Japan), IEEE Computer Soc., April 1998
9. De Silva, L.C., Ng, P. C. Bimodal emotion recognition. Automatic Face and Gesture Recognition, 2000.Proceedings. Fourth IEEE International Conference on, 28-30 March 2000. Pages: 332 – 335
10. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C., & Kazemzadeh, A. et al. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. Proceedings Of The 6Th International Conference On Multimodal Interfaces - ICMI '04. doi:10.1145/1027933.1027968
11. E. Mower, MJ. Mataric, and SS. Narayanan, “A framework for automatic human emotion classification using emotion profiles,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 19, no.5, pp. 1057–1070, 2011.
12. Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. Proceedings Of The 25Th International Conference On Neural Information Processing Systems - Volume 1, 1097-1105. Retrieved from https://dl.acm.org/citation.cfm?id=2999257
13. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780. doi:10.1162/neco.1997.9.8.1735
14. Deep learning for robust feature generation in audiovisual emotion recognition - IEEE Conference Publication.
15. Multimodal emotion recognition using deep learning architectures - IEEE Conference Publication.
16. Poria, S., Chaturvedi, I., Cambria, E., & Hussain, A. (2016, December). Convolutional MKL based multimodal emotion recognition and sentiment analysis. In Data Mining (ICDM), 2016 IEEE 16th International Conference on (pp. 439-448). IEEE.
17. Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile. Proceedings Of The International Conference On Multimedia - MM '10. doi:10.1145/1873951.1874246

18. V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad,"Ensemble of svm trees for multimodal emotion recognition," in Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific. IEEE, 2012, pp. 1–4
19. Vielzeuf, V., Pateux, S., & Jurie, F. (2017). Temporal Multimodal Fusion for Video Emotion Classification in the Wild. Arxiv.org. Retrieved 20 August 2018, from https://arxiv.org/abs/1709.07200
20. Tzirakis, P., Trigeorgis, G., Nicolaou, M., Schuller, B., & Zafeiriou, S. (2017). End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. IEEE Journal Of Selected Topics In Signal Processing, 11(8), 1301-1309. doi:10.1109/jstsp.2017.2764438
21. Huang, Z., Stasak, B., Dang, T., Wataraka Gamage, K., Le, P., Sethu, V., & Epps, J. (2016). Staircase Regression in OA RVM, Data Selection and Gender Dependency in AVEC 2016. Proceedings Of The 6Th International Workshop On Audio/Visual Emotion Challenge - AVEC '16. doi:10.1145/2988257.2988265
22. M.-J.Han, J.-H.Hsu, K.-T.Song, andF.-Y.Chang. Anewin- formation fusion method for bimodal robotic emotion recognition. JCP, 3:39–47, 2008
23. D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In The IEEE International Conference on Computer Vision (ICCV), December 2015.
24. Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The eNTERFACE’05 Audio-Visual Emotion Database. 22Nd International Conference On Data Engineering Workshops (ICDEW'06). doi:10.1109/icdew.2006.145