

Affective Computing
Dr. Abhinav Dhall
Department of Computer Science and Engineering
Indian Institute of Technology Ropar

Week - 08
Lecture - 02
Multimodal Affect Recognition

Hello and welcome. I am Abhinav Dhall from the Indian Institute of Technology Ropar. Friends, we are going to talk about Multimodal Emotion Recognition. This is the second lecture in the series for discussion on the topic of multimodal emotion recognition as part of the Affective Computing course.

(Refer Slide Time: 00:44)



Content



- Databases
- Benchmarks
- Multimodal Systems






So, here are the contents which we will be discussing. First, I would be bringing in front some of the databases, which have been proposed in the community and have been used for

analysis of emotion through multiple modalities. Then I would be talking about a very important aspect of the progress which is brought in through the benchmarks which are available in the community and later on some examples of the proposed multimodal systems.

So, the intent is to actually see how the information coming in from different modalities, different sensors that is fused. Now, just a recall so, in lecture 1 for multimodal emotion recognition, we discussed about why multimodal information is useful. To this end, we talked about how unique information can be brought into a system for analysis, when we use multiple sensors.


For example, we can combine a microphone with a camera. So, you have the voice and the face or you could have the text information along with physiological signals through EEG or ECG or text let us say with the face data. So, there is complimentary information which can be extracted from these modalities and we would like to learn from these different modalities so as to have a more accurate prediction of emotion recognition. Alright, let us dive in.

(Refer Slide Time: 02:22)


 **SEMAINE**  

- Sensitive Artificial Listener (SAL): A multimodal dialogue system with the social interaction skills needed for a sustained conversation with a human user.
- Aims to engage the user in a dialog and create an emotional workout by paying attention to the user's non-verbal expressions and reacting accordingly.
 - E.g. Nodding and smiling.

Affect sensing



Source: Douglas-Cowie et al., 2008

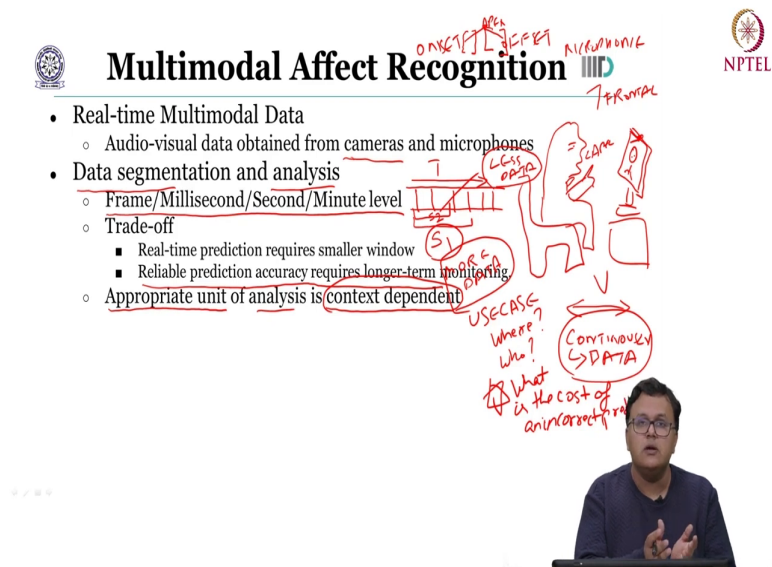


So, friends the first database which I would like to discuss with you is the SEMAINE dataset. Now, the SEMAINE dataset is a Sensitive Artificial Listener based dataset. What we have here is a multimodal dialogue system with social interaction skills, which are needed for a sustained conversation with a human user.

Yes, you would have guessed it right, as you can see here on the slide. Here you have these virtual avatars, virtual agents with which a user would be interacting and the intent is to drive a conversation so that emotion is elicited into the user. The aim of the SEMAINE dataset is to engage the user in a dialogue when conversing with these virtual agents and create an emotional workout by paying attention to the users non-verbal expressions and reacting accordingly.

Now, recall non-verbal expressions could include the facial expressions, your eye gaze and the body pose movement. And this is again as we have been talking about the affect sensing part. Once the system has sensed the affective state of the user then through these virtual avatars we are going to react, right. So, that is how the feedback to the user is going to go. Now, an example of this could be nodding and smiling you know these are non-verbal gestures.

(Refer Slide Time: 04:13)



Multimodal Affect Recognition

- Real-time Multimodal Data
 - Audio-visual data obtained from cameras and microphones
- Data segmentation and analysis
 - Frame/Millisecond/Second/Minute level
 - Trade-off
 - Real-time prediction requires smaller window
 - Reliable prediction accuracy requires longer-term monitoring
 - Appropriate unit of analysis is context dependent

Handwritten notes: ONSET, AFFECT, MICROPHONE, FRONTAL, LESS OPTIC, USE CASE? Where? Who? What? at the cost of accuracy?, CONTINUOUSLY DATA.

Now, further the aim is to have real-time multimodal data, ok. So, let us say the study is going on wherein you have this virtual avatar which is going to interact with the user. What we would have in front of the user is a set of cameras so that we can record the user. Now, again in this there are multiple aspects.

Let us say here is a person who is sitting on a chair ok, and here you have a computer screen. Now, you can have a camera right here let us say at the top of the monitor. This will give you the frontal view and you can also have other cameras to capture the user from different views. You can have a microphone let us say the lapel microphone which will get the clear sound from the user. And you can also have another microphone let us say the hanging microphone to record the ambient noise in the room.

So, this way you can have multiple sensors and get information about the conversation and the affect, which is elicited into the user. Further we want to do data segmentation and analysis. Now, imagine here you have the virtual character which is interacting with the user. So, we are continuously recording data, ok. So, this data is coming continuously. So, we need to segment divide the data into chunks so that those could be further used for assigning the labels.

And from the perspective of learning of a multimodal affect recognition system from the values which are going to be analyzed within a specific time duration. So, you can say well I am going to have a chunk which is segmented from a long video and this chunk would be of a certain specific duration and I am going to use this as one data point while I am going to learn a multimodal system.

Now, this segmentation is going to be at frame level and then further at you know the millisecond, second and minute level. There are certain trade-offs here. One could say well I have been recording this data let us say this is the timeline ok. So, this is the timeline this is T.

So, these are the different video frames which are being captured. I could take a window for segmentation let us say this much, ok. So, let us say this is S 1 which is segment 1. Another option is well, I could have taken from this frame to this frame only and I could have called this as segment S 2.

Now, obviously, in S 1 you have more data and in S 2 you have lesser data. But from a computation perspective this S 1 which contains more spatiotemporal data would require larger computation resources as compared to S 2 which has lesser data, right. So, the trade-off here is essentially what is the frequency at which we want to do the prediction of affect? Which is going to decide the duration of the chunk, which you are going to segment from the continuous data, which you are recording.

If you have a smaller window yes, it could have lesser temporal data, but we could do a more close to real-time analysis of the affect. However, if it is a longer window, you would require more computing resources then the output of the system might not be closer to real-time, but this is essentially dependent on the use case where you are deploying the system. So, the use case will decide the frequency at which we need to measure the effect of the user.

Now, another point to note in this trade-off of duration of the sample is, the reliable prediction accuracy may require longer term monitoring. Now, what does that mean? If you consider S 1 it contains more frames, more data then we can get extra information, more detailed information about the changes which are happening in let us say the facial expression of the user.

So, these changes which are happening over the time will give me important information and a more accurate prediction because I have more temporal data. However, if it is a shorter window, it is possible that the very few frames which are captured in this small window, I may not have got enough information to actually predict the correct affect. So, what that could mean is let us say a person starts to smile.

Now, they are starting to smile so there is an onset of an expression, typically it has to go to the apex of the expression the highest intensity and then it might go down to offset. Let us say the person who is going to smile, but what happened is we were considering a smaller window.

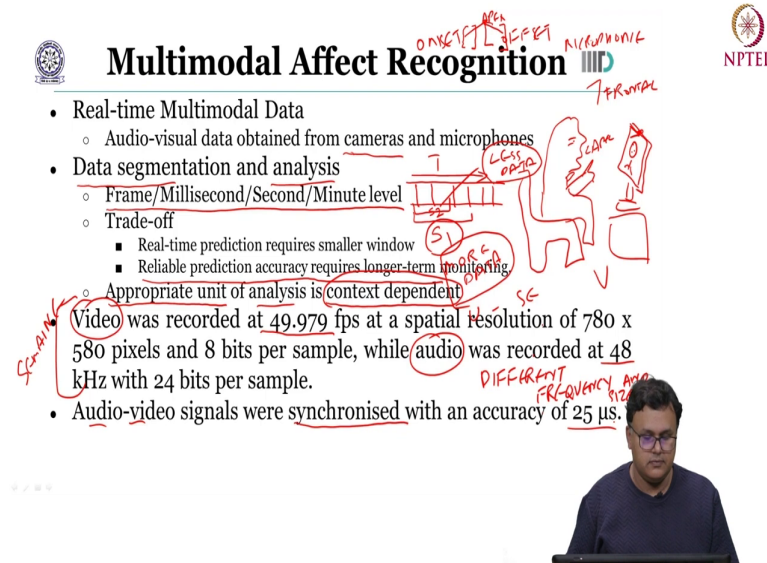
So, we only had the data from onset to let us say midway to apex. Now, this might not be enough for the system to be robustly able to predict that the person is smiling, it could let us say very well be confused with that that the person was just speaking. So, lip close, lip open.

Now, if we have a longer window let us say I am now going to focus on apex to offset all of the frames here. Now, this could give me a more reliable information with the extra frames, which now I have to analyze. So, I can see the transition between the facial expressions which can give me a more robust information.

Further, it is extremely important to have the appropriate unit of analysis. And what we understand is that the appropriate unit of analysis you know how long you want to have, how many frames, what is the duration that is all context dependent that is the use case dependent, right.

It is about where are you going to use the system and who is going to use the system and an extremely important point here friends what is the cost of a incorrect prediction. So, an example is let us say we are going to use a multimodal affect recognition system in a health and well being based use case. So, the cost of doing an incorrect prediction is far higher as compared to let us say another scenario where we are going to use a multimodal affect recognition system to suggest movies to a user.

(Refer Slide Time: 12:32)



The slide is titled "Multimodal Affect Recognition" and features the NPTEL logo in the top right corner. It contains a bulleted list of topics related to real-time multimodal data and data segmentation. Handwritten red notes and a diagram are present. The notes include "ONCE IT IS DONE", "MICROPHONE", "7 frames", "LESS DATA", "SCAN", "SYNCHRONISATION", "DIFFERENT FREQUENCY AND SIZE", and "SCAN". The diagram shows a person's head and shoulders, with a box labeled "SCAN" and a box labeled "SYNCHRONISATION".

- Real-time Multimodal Data
 - Audio-visual data obtained from cameras and microphones
- Data segmentation and analysis
 - Frame/Millisecond/Second/Minute level
 - Trade-off
 - Real-time prediction requires smaller window
 - Reliable prediction accuracy requires longer-term monitoring
 - Appropriate unit of analysis is context dependent


Video was recorded at 49.979 fps at a spatial resolution of 780 x 580 pixels and 8 bits per sample, while audio was recorded at 48 kHz with 24 bits per sample.

Audio-video signals were synchronised with an accuracy of 25 μ s.

Now, from the SEMAINE perspective this is the data set. The video was recorded at 49.97 frames per second at a spatial resolution of 780 cross 580 and the researchers they kept the bit rate of the data at 8 bits per second, while the audio was recorded at 48 kilohertz with a 24 bit per sample. Now, what that mean friends here is you have the video and then you have the audio, which have different frequency and size of data, right.



Therefore, synchronisation is required as we have discussed in the earlier lecture. Now, in the case of SEMAINE the researchers they synchronised the audio and video with an accuracy of 25 milliseconds. Now, let us look at other attributes of the SEMAINE dataset. We are capturing spontaneous data.

(Refer Slide Time: 13:40)



SEMAINE Database

- Spontaneous data capturing the audiovisual interaction between a human and an operator undertaking the role of an avatar with four personalities:
 - Prudence, who is even tempered and sensible; *category AC*
 - Poppy, who is happy and outgoing;
 - Spike, who is angry and confrontational; and
 - Obadiah, who is sad and depressive. *VA 2 dimensions*
- All interactions were annotated by 2 to 8 raters in the four dimensions in continuous time and continuous value.
- Dimensions are Arousal, Expectation, Power, and Valence.
 - Four dimensions account for most of the distinctions between everyday emotion categories (Fontaine, J., et. al., 2007).
- $\{v_i^a, v_i^e, v_i^p, v_i^v\}$ for every rater i and dimension $a/e/p/v$.



Now, this is based on the audio-visual interaction between the human and the operator which is undertaking the role of an avatar with four personalities. So, the virtual avatar can have four different type of personalities. Why do we want this? Well, if the avatar is going to have different personality, then this can affect the course of conversation between the user and the avatar, right.

In simpler words, if let us say two people are talking, the direction of the conversation would be affected by the personality of the individuals and that is what the researchers also wanted to capture. So, that you know we have spontaneous data and along with a spontaneous data we can also reflect on the nonverbal behaviour which would be induced into the user due to the personality of the virtual avatar.

So, the four personalities are as follows. The first is Prudence. Now, this virtual avatar is even tempered and sensible. The second is a Poppy virtual avatar who is happy and outgoing. The third is Spike. Essentially, this is angry and confrontational and the fourth is Obadiah who is sad and depressive. Now, you can very well imagine in these four cases, the response of the avatar would be reflective of the affective state of the avatar and the personality, right.

So, if you have a happy outgoing virtual avatar. So, it is possible that the affect which is induced into the user who is interacting with the Poppy, happy avatar could also be a slightly more on the positive affect. As compared to let us say when the same user is interacting with the fourth one, Obadiah where the virtual avatar's personality is a bit sad, right.

So, the response from the user against the conversation which is going to happen with this fourth avatar could be a bit neutral or could be a bit negative, right. So, the personality is going to affect the conversation. Further guys in SEMAINE, all interactions they were annotated by 2 to 8 raters in four dimensions in continuous time and continuous value. Now, these four dimensions are arousal expectation, power and valence.

Typically, when we were discussing about the emotion analysis from faces earlier or emotion analysis or from voice, we were talking about either the categorical emotions or we were talking about valence and arousal two dimensions only. In this case, the authors they also got labelled expectation and power as well. Now, the rationale is as follows, if you have four dimensions, they account for most of the distinctions between everyday emotion categories.

So, you can have fine-grained emotion labels when you are having four dimensions to represent the emotion. Now, here $V_a^{(i)}$, $V_e^{(i)}$ and p_i here are indicating the ratings which was given to a particular sample by rater i and the dimensions in the superscript are representing the four dimensions of the emotion.

(Refer Slide Time: 18:01)

Emotion Recognition in the Wild

- EmotiW challenge series
- Audio-visual emotion recognition task (Dhall et al., 2018)
- Acted facial expressions in the wild
- Labels early initialized through closed caption parsing


Handwritten notes:
Challenging conditions
Representative of Real-world scenarios
Database → Parsing of captions

Now, guys let us move to another benchmark which is very commonly used in the affective computing community. Now, this benchmark is again based on a series of challenges which are hosted by researchers. So, this one is called the emotion recognition in the wild. And recall in the wild here simply means you have challenging conditions which are representative of real world scenarios.



Now, let us discuss about EmotiW. So, EmotiW has several task for affect analysis. We are going to talk about the audio visual emotion recognition task. Now, this is based on the acted facial expressions in the wild data base, which we collected from high quality Hollywood movies which is based on parsing of captions.

So, essentially captions which are available for viewers with hearing disability. So, we pass those captions, we got those sample videos which could contain a user or group of users showing some emotion and then the labellers they had the final label for each video.

(Refer Slide Time: 19:57)



Emotion Recognition in the Wild

□ Performance comparison of different methods on EmotiW audio-visual emotion recognition task.

① Data is limited

② Varied environment videos

intra class variability

Data point + Meta-data

Category emotion

Metric - evaluation classification accuracy

2018

Rank	Team	Class Accuracy (%)
1	SinoTech [1]	61.59
2	E-Emotion [6]	61.19
3	AUTP [12]	60.64
5	GU-UC [17]	60.04
6	UoT	60.00
8	NLPB	60.34
9	INRIA	59.72
10	SUAT	59.04
11	Tsinghua University	57.12
19	AUT-LAB	56.51
21	VIT	56.05
12	UGBC	55.74
13	VIPL-ACF-CAS	55.59
14	Imag	55.13
15	ZINC Lab	54.98
16	Summerlings	54.82
17	EmotLab	54.21
18	CNU	53.75
19	Mind	53.60
20	Midea	53.45
21	Korea University	53.14
22	Kathu	51.76
23	Beijing Normal University	50.54
24	USTC, NUS&UP	49.76
25	PopView	48.09
26	BUCT	45.94
17	BBCLab	43.57
28	CofredLab	41.81
29	Baseline	41.07
30	12-4C	35.89
31	SAANWILD	33.84

Source: Dhall et al., 2018

Now, the EmotiW challenge based on the audio-visual emotion recognition task has seen a healthy representation and participation from both academic and industrial labs. Now, the task is you get one data point, data point is audio visual and you predict the category of emotion, ok.



What is also available along with this audio visual data is Meta data, which tells things such as the age of the subjects in the video or the gender. Now, the metric of evaluation is

classification accuracy. So, here you see the classification accuracy which is a comparative of the different methods which were proposed in 2018. So, this is from 2018.

Now, the point to note here friends is the highest performance is not really high. So, this was a categorical emotion task. Still, we are around 61 percent classification accuracy. Now, there are several reasons. We have a multimodal emotion recognition system. So, all these teams they propose different multimodal emotion recognition systems, analyzing audio and video some teams also added text analysis by doing speech to text from the voice part of the video.

But the reason essentially for this performance is that data is limited around 4000 videos. Now, given the very varied environments you know which are reflected in the videos because the videos in the data set are collected from Hollywood movies, the limitation of the data and the intra class variability introduced due to the different environments that makes the task non-trivial.

(Refer Slide Time: 22:42)




 **Audio-Visual Emotion Challenge** 

Uni and multi-modal challenge based on acoustic, linguistic and video cues

arousal expectancy power valence

low



high




Now, similar to the emotion recognition in the wild challenge is a very successful and extremely well used benchmarking effort called the audio visual emotion challenge. Now, in the audio visual emotion challenge there have been several task similar to EmotiW. Now, here is one such example where friends you can see the four different dimensions of continuous emotion and the low and high intensities for the same.

You can actually make out a difference the visible difference between the facial expression well, let us say the arousal is low or arousal is high, right. Now, the I have a challenge the audio visual emotion challenge not only had the multi-modal task audio visual task, but uni-modal task as well, which were based on the acoustics, linguistic and the visual cues coming from the data.


(Refer Slide Time: 23:42)



Data and Annotations



- To attain binary labels,
 - Average value of each dimension over all raters was computed, resulting in a set of continuous time, real valued variables $\{v^e, v^c, v^p, v^v\}$
 - Mean of these average ratings over all interactions in the dataset was computed, resulting in the scalar values: $\{\mu^a, \mu^c, \mu^p, \mu^v\}$
 - The binary labels $\{y^a, y^c, y^p, y^v\} \in \{\pm 1\}$ are then found by thresholding $v_j^i > \mu^j$ for each dimension j at every frame t .
- Video stream was chunked in two ways:
 - Frame level: Per frame for the video only tasks
 - Word level: Per word for the audio and audio-visual tasks

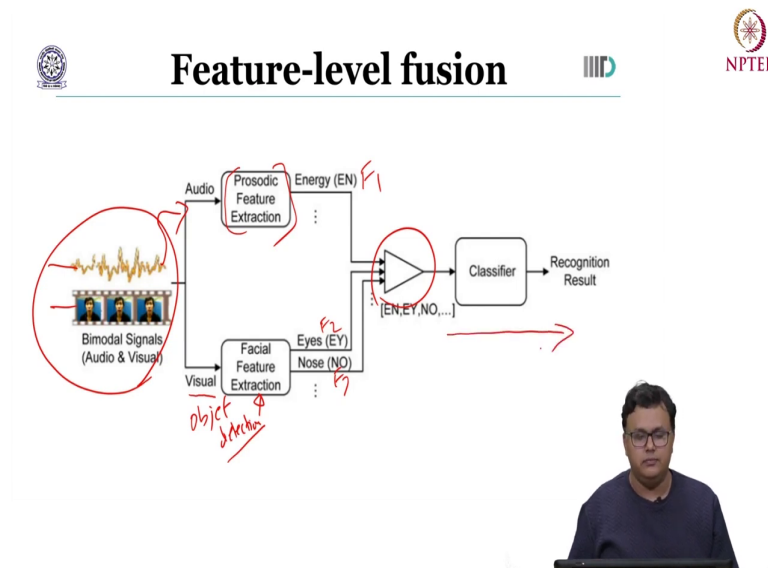


Now, in order to obtain the binary levels for the average data the average value of each dimension over all datas was computed, ok. So, this is again coming from the SEMAINE data which resulted into continuous time real variables which are V^a, V^e, V^p and V^v and again guys these are representing the emotion dimension.

Now, further to get the binary labels mean of the average ratings over all interactions in the dataset was computed which resulted into these scalar values. Now, if you recall from the earlier slide, you had low and high binary labels, right. These binary labels were essentially based on thresholding the values which were based on the mean of all the labellers who labelled a particular video sample.

Further for AVEC the video streams they were segmented in two ways. First is frame level. Now, you would like to analyze per frame for the video only tasks. The second is you there was a word level segmentation so, per word for the audio and audio-visual tasks.

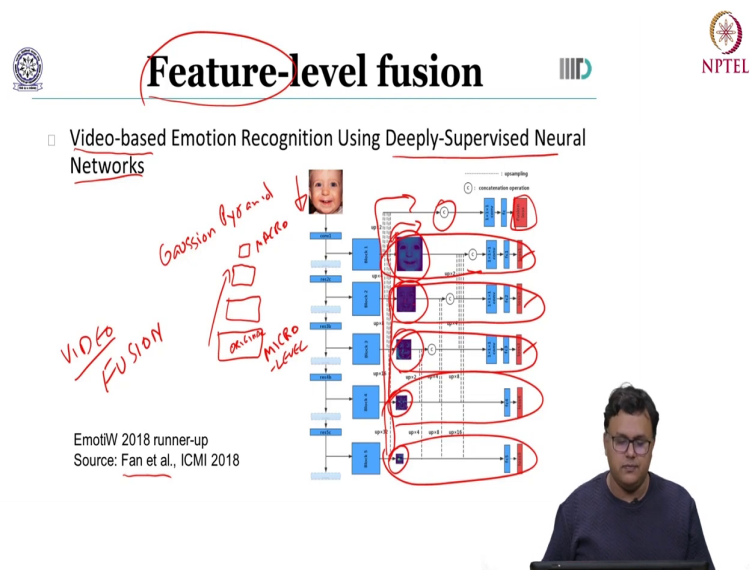
(Refer Slide Time: 25:09)



Now, here is an example of the feature level fusion based task for AVEC. So, friends what do we see here? We have an audio-visual signal coming in, the video part the audio part. Now, the audio part that is input into a library to compute the prosodic features and in parallel the video part has object detection to detect the location of the face and then from that the features are extracted.

Further the features which are coming let us say F 1 and F 2 for eyes and different parts they are all fused together. So, here you have feature fusion and then there is a classifier to predict the binary label.

(Refer Slide Time: 26:01)



Now, what we are going to do from this point is we are going to look at different works which are proposed in the literature and we are going to see how researchers proposed these multi-modal affect recognition systems which are combining information at different level. Now, the first one for feature level fusion friends I would like to discuss with you is the work called video based emotion recognition using deeply supervised neural networks.

So, this is actually a work from Fan and others in 2018. Now, notice there is only one modality in this work that is video. However, there is fusion happening at multiple levels how

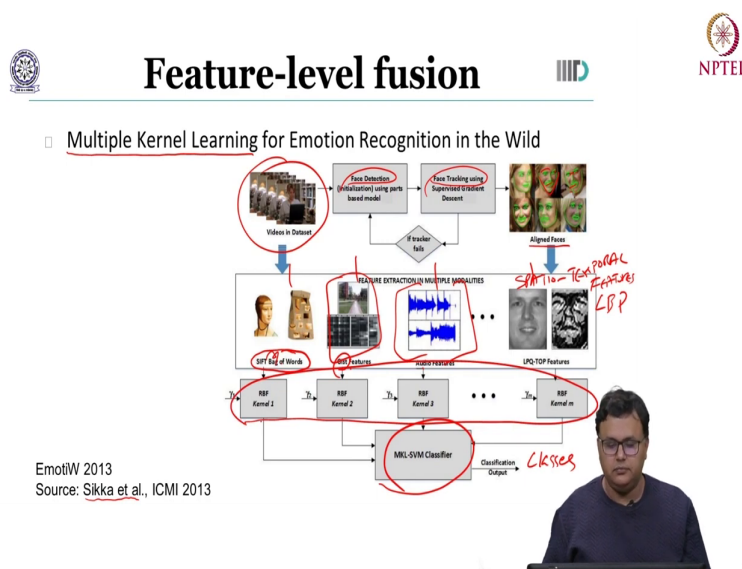
is that? So, here you have an input face. Now, the architecture is going to have the face represented at different resolutions or different scales.

So, essentially, we are going to mimic a Gaussian Pyramid where you say well, I have the smallest resolution of the image 1, a larger resolution version of that image then further the larger image and then the largest. The aim is this was the original image which I had got, right. So, I was down sampling I was going up. Now, at the bottom of this pyramid I would be able to analyze micro level information from the perspective of face.

Think of it as the twitch, subtle twitch in the corner of the eye. As we go up this pyramid, we have the macro level information which will give you let us say the just the face, tens of face and if it was smiling a stretch of the region in the lower half of the face. Now, the authors they introduce the concept of scale within the network itself wherein when the input face comes in you have a series of configuration layers then you have these down sampled versions of the feature maps.

Now, what we are doing here is, we are actually learning to predict the emotion from each scale feature maps individually. So, you can see the loss here and in parallel we are also combining the feature maps. So, all of these go together and they are fused. So, this is the concatenation which is happening ok, and this is the final fusion loss. So, essentially first you fine tune, you have discriminative we presented the features at each level and then you are doing the fusion.

(Refer Slide Time: 29:02)



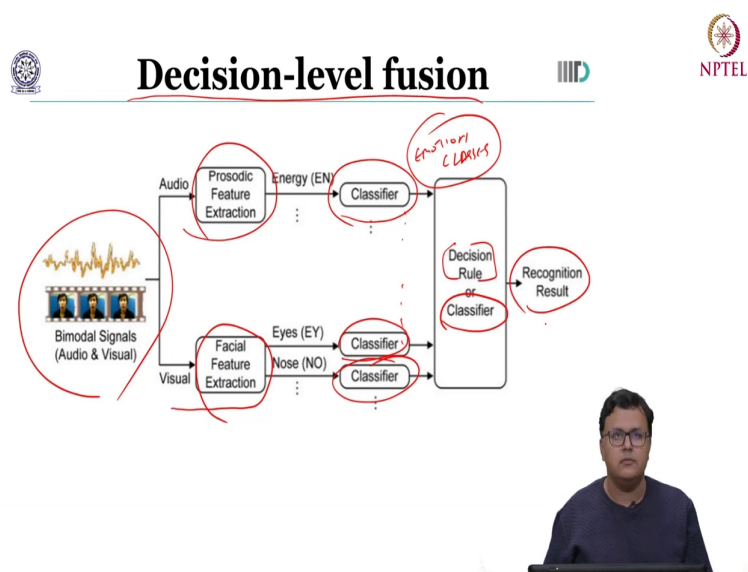
Now, friends let us move to another work. So, this is the non deep learning based work by Sikka and others. So, this is called multiple kernel learning for emotion recognition. Now, let us look at how information is fused in this part. Here you have the input audio-visual sample we do object detection for getting the face, we track the facial points. So, here you can see the facial points.

So, as you have already seen in the tutorial for the automatic facial expression recognition there is this several open source libraries for object detection and facial pass detection. So, you can use that, authors used one such library and then they have the aligned faces as input. What they are doing is they are computing a scale invariant feature transform a hand engineered featured feature in a bag of words based framework. We have discussed this in automatic facial expression recognition.

Then they want to analyze the whole scene. So, this is another hand engineered feature called gist. Along with that they have the audio feature, ok. So, they are extracting audio features and then they also have spatiotemporal features which are inspired from your local binary patterns. Now, what do we have here? We have these set of features audio-visual features which are extracting different type of information.

So, to combine them the authors they would apply different kernels to each of the individual features and then learn together a multiple kernel learning based support vector machine to predict the emotion classes. Now, in this case the fusion essentially happens at the kernel level. Now, these were some examples of early fusion.

(Refer Slide Time: 31:01)

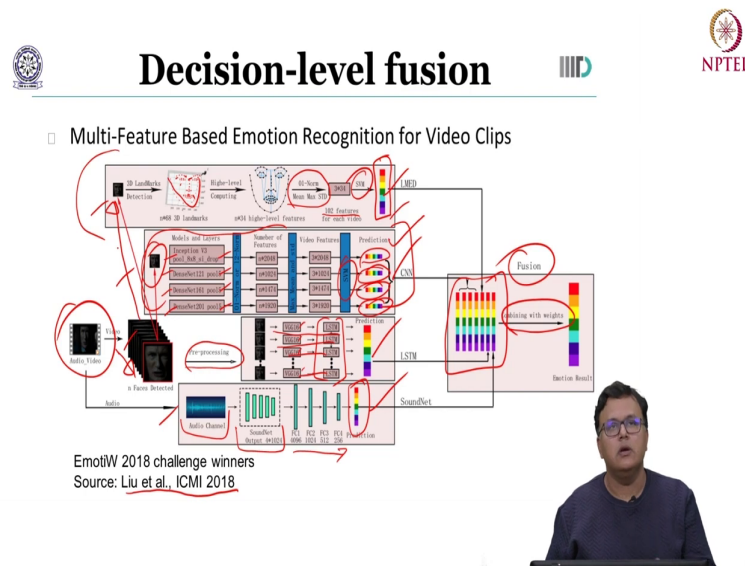


Let us look at some examples of late fusion. You have again the audio-visual signal coming in, you extract some audio features, you have a classifier here could be a support vector

machine your hidden Markov model, which gives you the emotion classes in the form of probabilities. In parallel you have the visual channel.

Similarly, you extract the features, then you have these classifiers again these give you the emotion classes. Now, you can do two things you can have decision rules for example, applying and or kind of gates to the different classifier outputs or you can have a classifier. So, you can combine the output of all these classifiers and learn a classifier on top of it to get the emotion class.

(Refer Slide Time: 31:57)



Now, the first work in this which I would like to bring to your attention is from Liu and others from 2018. Now, let us see how the fusion is done, ok. So, you have a audio-visual sample again let us look at the first video you do object detection, what the authors do is they

first look at the facial landmark points, which tells you the location similar to how we saw in the earlier works.

And then they are computing the mean max standard deviation for all the frames and making a 102 feature dimension for each video. Then they use a support vector machine and here is their emotion classification. So, these are the classification results for emotion when the video is analyzed based on the facial points.


The second is we take the face we then have different pre-trained networks. So, these are pre-trained convolutional neural networks, we extract the features and then we are learning a support vector machine individually for each of them.

Again, what do you see friends? Series of predictions of emotion based on these classes. Now, the third is you take the faces you input into a VGG pre-trained network. So, again a face based communicational neural network and then you have a recurrent neural network for each of them separately which gives you the predictions.


Now, let us look at the audio, you have the audio channel the pre-trained sound net is used, you have a series of fully connected layers, you get the predictions. So, look at the levels at which we have got the emotion classes. Now, what we are we doing? We are combining them together into a grid and then we are combining them with weights. So, this is the decision fusion part.

So, we optimized the prediction for each feature, which we are extracting here and once we got the predictions, we are then fusing them with different weights. So, with empirical evaluation, with experimentation researchers realized for example, this performs better as compared to this. So, let me give more weightage to this pipeline as compared to this pipeline and then you can do the combination with weights.

(Refer Slide Time: 34:31)


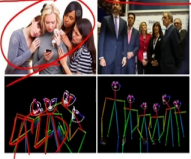


Decision-level fusion




- Group-Level Emotion Recognition using Hybrid Deep Models based on Faces, Scenes, Skeletons and Visual Attentions

Context Attributes scene & subjects



EmotiW 2018
Source: Guo et al., ICMI 2018

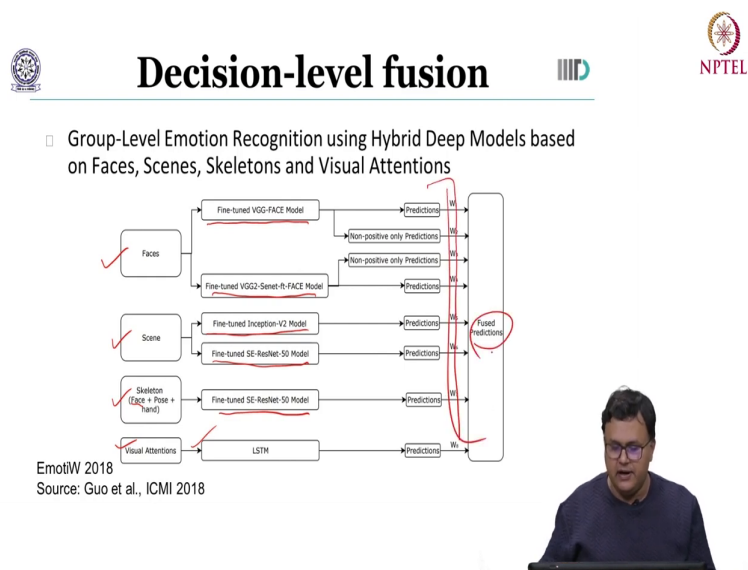


Now, here is another work for fusion. You notice here this is group level emotion task. So, now we are going to talk about when in a given sample you have multiple people. Now, this work is from Guo and others. And the authors they propose the use of face level information, scene level information, the body pose in the form of skeleton and the visual attention. So, let us say here you have a sample frame this is the input frame.

What you see here is the facial structure. So, this is the facial structure and the lines here are telling us about the body pose, right. So, along with this the user is also analyzed by using object detector. So, you can see here for example, there is a tall building in the back and then this person is wearing a black pant blue jeans and so forth. So, these are the visual attention based attributes about the scene and the subjects.

So, this gives you the information about the context right, where people are who these people could be nor from the identity perspective. But from the perspective of attributes such as young people or aged people could be school going students or could be colleagues in an office, right.

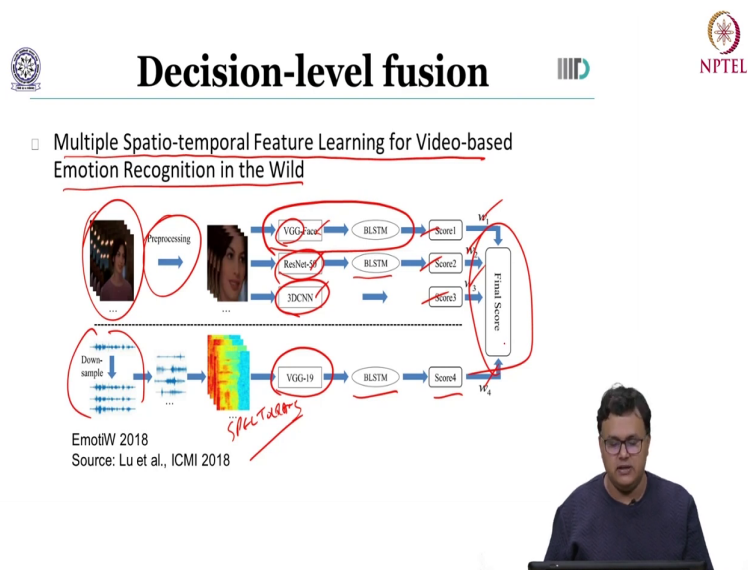
(Refer Slide Time: 36:07)



So, face detection is used you get the faces, scene is the whole image, then you have the body pose and the face structure and visual attention. So, these are the input you have the VGG pre-trained face model which is used to analyze the faces, then you have the scene based model which is looking at the whole image to get the different attributes.

Further the skeleton body pose, arm up, arm down all this is analyzed and then the visual attention again this gives you the attributes. Later the predictions are fused here to get the final result.

(Refer Slide Time: 36:51)

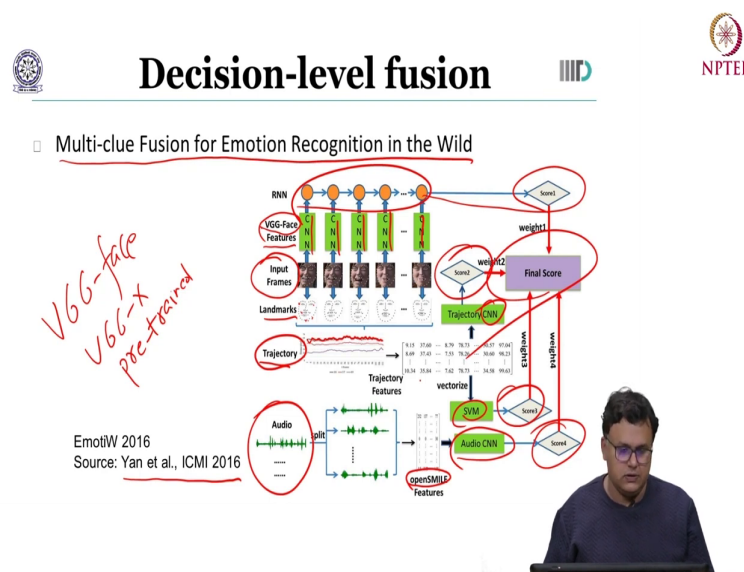


Now, let us look at another work for fusion. Friends, this is multiple spatiotemporal feature learning for video based emotion recognition in the wild. In this case, here you have the faces, we do some pre-processing and you see three different type of networks.

Now, notice this is again a recurrent neural network where the feature to each cell is the a feature which is extracted from a pre-trained VGG network, then you have a ResNet based network, again a recurrent neural network, this is the bidirectional LSTM and here the authors have a 3D convolution neural network, ok.

Now, you get the scores from each of them and we are going to fuse in parallel, we have the audio information, we get the spectrograms, we input the spectrogram into the pretrained face VGG network, we have bidirectional STM, we get the scores, we do the final fusion, ok. And here, W_1 , W_2 , W_3 , W_4 are signifying the weights for these individual feature based scores which we get from the classifiers of these features.

(Refer Slide Time: 38:05)



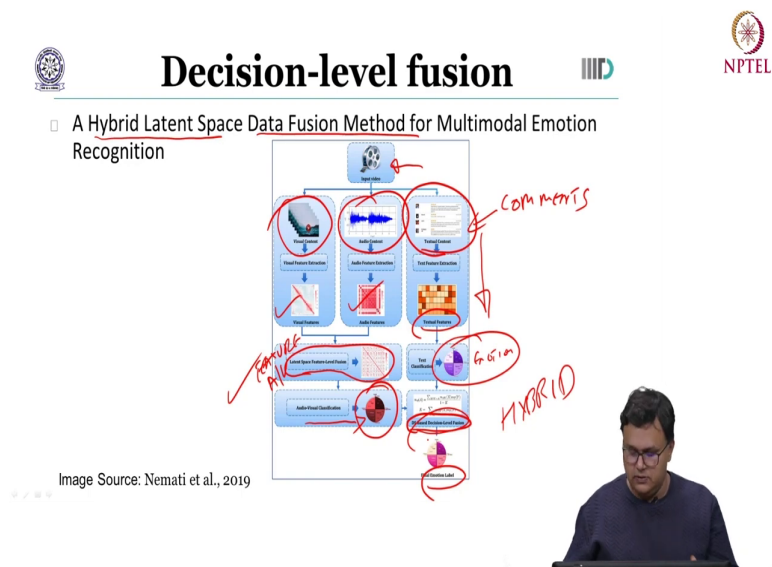
Now, here is another work from 2016 friends called multi-clue fusion for emotion recognition in the wild. Now, let us see what is proposed in this case. Here you have the input video and the input audio, you again detect the landmarks, you get the facial points where are the different facial paths, you create trajectories of these landmarks, ok you can. So, you can create how are the facial point changing over time, right. You input that into a computational neural network.

Separately, you also again extract VGG face. Now, you would have noticed VGG face and VGG based pre-trained network that is very commonly used in the community for face analysis, simply because these network especially VGG face that has been trained on millions of face images. So, it has very rich representation with respect to the different attributes of faces.

Now, coming back we have the VGG features, we are extracting these features separately for each frame, then we have a recurrent neural network, ok. Now, this gives you the scores here. Similarly, we already had the scores for the trajectories. Now, coming to the audio, you have open smile with features.

Now, open smile again is a very commonly used library to extract the features, you input that into a audio computational neural network and what you are doing is you get the scores. These scores and then another score from trajectories by using support vector machine and then you are fusing them to get the final emotion class.

(Refer Slide Time: 39:55)



Now, here is another work friends by Memati and others. This is called a hybrid latent space data fusion method for multimodal emotion recognition. Let us go through this, you have the input sample, you extract the visual features, you extract the audio features. What you also do is you have some textual data. Now, if you notice these are some comments about the video.

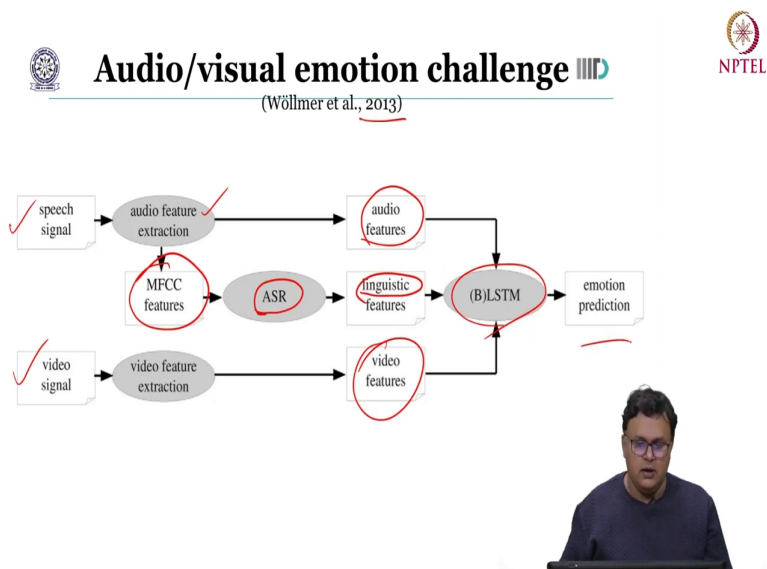
Now, you extract visual features, audio features and you also extract textual features. Now, notice the different level of fusion which is happening. What the authors propose is first we do a fusion of the audio-visual data and then we have a common classifier for predicting the emotion from the audio-visual fusion only. So, this is like a feature level fusion which happens first, ok.

So, notice this is feature level for audio and video. Separately for the text data, we do the text classification and get the emotion. Now, we have emotion information from this part,

emotion information on this part. What we are doing is we are doing decision level fusion; you know a late fusion to predict the final emotion label. Now, interesting thing to note here is the audio-visual data is more similar in terms of frequency as compared to audio and text or video and text.

Even the audio-visual data is very different, but relatively, right. So, the authors when they experimented with the different combinations, here they actually come up with a hybrid solution, right. For some part of the features, they have feature level fusion, they get the prediction from that and then they fuse at the decision level after they get the scores from all the different features.

(Refer Slide Time: 41:56)



Now, let us look at examples from the audio-visual emotion challenge friends. So, in this case, if you look at Wollmer work in 2013, very similar to the pipelines which we have

discussed till now, you have speed signals, you have video signals, you extract your standard MFCC feature and what you are doing here is you are doing automatic speech recognition.

Now, this will give you some linguistic features, the content, what is being spoken and from the audio features, you will get the rich features for example, again you know MFCC, your fundamental frequency, intensity and so forth. For the video, you extract video features and then you are combining them together into a bidirectional LSTM to predict the final emotion.

(Refer Slide Time: 42:48)

Results: No Feature Selection

Classifier	Features	AROUSAL		EXPECTATION		POWER		VALENCE		Mean
		WA	UA	WA	UA	WA	UA	WA	UA	WA
BLSTM	A	68.5	69.3	64.3	53.5	66.1	53.3	66.3	56.1	66.3
BLSTM	A+L	67.8	69.0	64.8	52.0	65.5	53.9	66.3	56.2	66.1
LSTM	A	68.5	68.6	66.1	55.9	64.7	56.1	65.6	55.2	66.2
LSTM	A+L	68.2	68.8	65.2	51.9	66.2	55.0	63.8	55.9	65.9
SVM [6]	A	63.7	64.0	63.2	52.7	65.6	55.8	58.1	52.9	62.7
BLSTM	V	62.3	62.9	62.3	51.8	55.2	53.0	63.3	60.5	60.8
LSTM	V	60.3	61.3	60.4	57.7	57.0	50.4	64.0	57.9	60.4
SVM [6]	V	60.2	57.9	58.3	56.7	56.0	52.8	63.6	60.9	59.5
BLSTM	A+V	67.7	68.0	63.1	53.4	60.6	55.0	67.2	61.8	64.7
BLSTM	A+L+V	66.9	67.0	66.2	57.3	63.4	52.3	65.9	61.5	65.6
LSTM	A+V	68.0	67.5	65.7	57.7	63.8	54.7	65.5	59.5	65.8
LSTM	A+L+V	67.4	66.8	65.3	56.7	61.7	54.2	67.6	62.8	65.5
BLSTM (LF)	A+V	67.9	69.3	65.0	53.2	64.0	55.5	69.8	61.3	66.7
BLSTM (LF)	A+L+V	67.0	68.6	65.7	51.6	63.6	55.7	69.8	61.2	66.5
LSTM (LF)	A+V	62.6	64.3	67.6	57.6	65.1	56.0	68.2	57.7	65.9
LSTM (LF)	A+L+V	66.3	67.4	63.9	58.1	66.0	53.9	66.4	58.2	65.7



Now, the authors did a very comprehensive analysis within this framework to see how useful feature selection is going to be when we have these different modalities coming in. Now, here friends you see the different classifiers, here we have the four dimensions for continuous emotion and here is the mean result, ok.

Now, for example, notice here the weighted accuracy for LSTM based classifier when the feature is only audio is highest for arousal. However, let me bring the discussion to the bottom part of the table here. When you have again your LSTMs, here you are not doing any feature selection, you see when the combinations are there for the audio video, audio linguistic and video, the performance is the highest.

So, what do we understand? Feature fusion is helping. And even though we are not selecting discriminative features, even then we see an increase in performance. However, recall that one of the challenges in creating a multimodal emotion recognition system is that when we have data coming in from different modalities, we would like to extract the unique information from these modalities, which can be complementary to each other so that we can do a more robust prediction.


If we have overlapping information, then we can have a long dimension feature, let us say if it was feature fusion which could have a lot of overlap, but unnecessarily we may be going into the course of dimensionality.

(Refer Slide Time: 44:39)



Discussion: No FS

- For arousal, the best WA of 68.5 % is obtained for acoustic features only, showing that audio is the most important modality for assessing arousal (M. Wollmer, et. al., 2010).
- The classification of expectation seems to benefit from including visual information as the best WA (67.6 %).
- Power is best classified via speech-based features. However, for unidirectional modeling WA significantly increases from 64.7 % to 66.2 % when using linguistics in addition to audio features
- For valence, the inclusion of video information helps, leading to a WA of 69.8 %.



Now, in this interesting work from AVEC, when no feature selection was used, we observed weighted accuracy was highest for acoustic features only. This also shows audio is the most important modality for arousal. Further, the classification of expectation dimension, it seems to benefit from including the visual information. So, when we do the fusion. For the power emotion dimension, it is best classified by speech based features.

Further, the authors noticed that unidirectional modeling of the weighted accuracy; it significantly increases when this linguistic features are also used along with audio. So, the fusion here is not only about the audio features, but also what is being said, the contextual information which you are getting from linguistics. For the valence emotion dimension, when we include video, that increases the weighted accuracy.

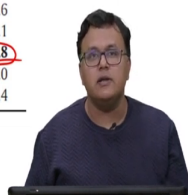
(Refer Slide Time: 45:36)



Results: CFS Feature Selection



Classifier	Features	AROUSAL		EXPECTATION		POWER		VALENCE		Mean
		WA	UA	WA	UA	WA	UA	WA	UA	
BLSTM	A	71.3	70.2	66.2	51.0	66.0	56.4	65.9	60.6	67.4
BLSTM	A+L	73.7	74.4	66.1	53.1	64.6	55.7	65.8	57.2	67.6
LSTM	A	70.4	69.8	67.7	54.6	64.9	58.8	63.1	55.3	66.5
LSTM	A+L	71.9	71.1	63.1	55.5	66.6	56.3	64.7	56.9	66.6
BLSTM	V	59.8	58.8	66.2	50.1	64.1	57.5	63.3	56.0	63.4
LSTM	V	62.7	61.5	66.0	50.1	70.2	62.4	64.3	52.7	65.8
BLSTM	A+V	67.8	69.5	64.3	52.3	60.1	57.0	64.7	58.8	64.2
BLSTM	A+L+V	69.9	70.7	63.3	50.4	61.9	56.1	61.4	55.9	64.1
LSTM	A+V	69.7	70.8	64.5	52.0	63.5	56.8	62.4	53.0	65.0
LSTM	A+L+V	70.4	71.3	65.7	53.3	63.5	55.9	62.9	53.2	65.6
BLSTM (LF)	A+V	68.5	67.5	66.7	50.4	64.2	52.7	69.1	60.6	67.1
BLSTM (LF)	A+L+V	72.3	72.3	66.6	50.9	64.4	54.0	67.9	58.5	67.8
LSTM (LF)	A+V	65.7	63.7	67.4	52.1	68.0	58.6	66.8	54.8	67.0
LSTM (LF)	A+L+V	64.8	63.5	67.1	54.9	68.1	57.3	65.7	56.4	66.4



Now, when feature selection is used, notice how the performances have gone up. When you have a bi-directional STM, this is audio plus linguistics, notice how now we are into the 70s with respect to the weighted accuracies and the unweighted accuracies, right.

So, the same is observed for when we have the fusion for audio linguistics and video, the performance actually goes considerably up. So, that means, feature selection is an important step when we are going to do fusion of data coming in from different modalities.

(Refer Slide Time: 46:14)



Discussion: FS



- For most settings, CFS does not significantly improve the average weighted accuracy.
- However, for recognition based on video only, CFS leads to a remarkable performance gain, increasing the mean WA from 60.4 % to 65.8 % for unidirectional LSTM networks.

4 / 44



Now, some other points with respect to the feature selection, it was noted that for most settings, it did not improve the average weighted accuracy, ok. However, for recognition on video only, it leads to a remarkable performance gain, increasing the mean weighted accuracy by 5.4 percent.

(Refer Slide Time: 46:38)



Discussion: AVEC Challenge



- Audio features lead to the best result for arousal classification.
- For classification of expectation using facial movement features, the obtained WA of 68.6 % is higher than what is reported for other techniques.
- For power, audiovisual classification with Latent-Dynamic Conditional Random Fields as proposed in (G. Ramirez, et. al., 2011) outperformed.
- For valence, the audio features lead to the highest accuracy (70.2 %) (A. Sayedelahl, et. al., 2011).

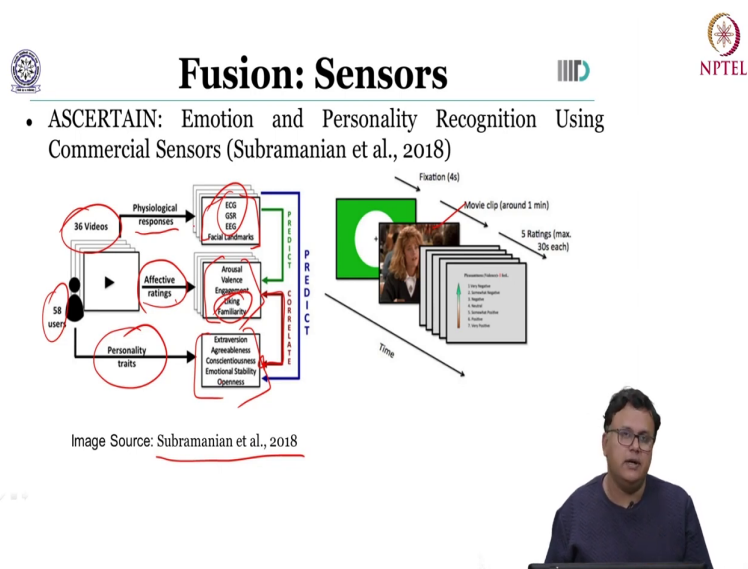
*AVEC
to
Benchmarking*



Now, some other interesting discussion takeaway points from the audio-visual emotion challenge benchmarking friends. Audio features lead to the best results for arousal classification. We have seen that across different works. For classification of expectation, facial movement features were very useful. For power, emotion dimension, audio visual classification with latent dynamic conditional random fields, which was proposed in 2011 by Ramirez and others, that was the state of the art in 2011.

For the valence, the audio features lead to the highest accuracy. Now, if you notice, these are the findings from the different works, because AVEC is a benchmarking effort, which has different tasks. So, all these are different papers, different methods, which were proposed through the analysis on the rich data, which is part of AVEC.

(Refer Slide Time: 47:44)






Now, I would like to change the discussion a bit and talk about methods, where we are doing fusion for with sensors, right. Till now, we have been talking about audio and visual data primarily. Now, here is a work from Subramanian and others. Now, this is essentially the researchers were interested in emotion and personality, ok.

So, they collected a very rich data, where they had a series of physiological sensors, which were attached to the users who were looking at videos. So, here you have 58 users, who looked at 36 videos each, physiological data was recorded. They gave affective rating to the videos and their personality traits were also evaluated.

Now, what the researchers did, they computed the correlation between the personality traits and the emotion, along with two other attributes, which are liking and familiarity. Now, they

noticed that you could actually predict the personality and the affect by looking at physiological sensors, by doing the fusion of the data coming in from these sensors.

(Refer Slide Time: 49:18)

 **Fusion: Sensors**  

- K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations (Park et al., 2020)







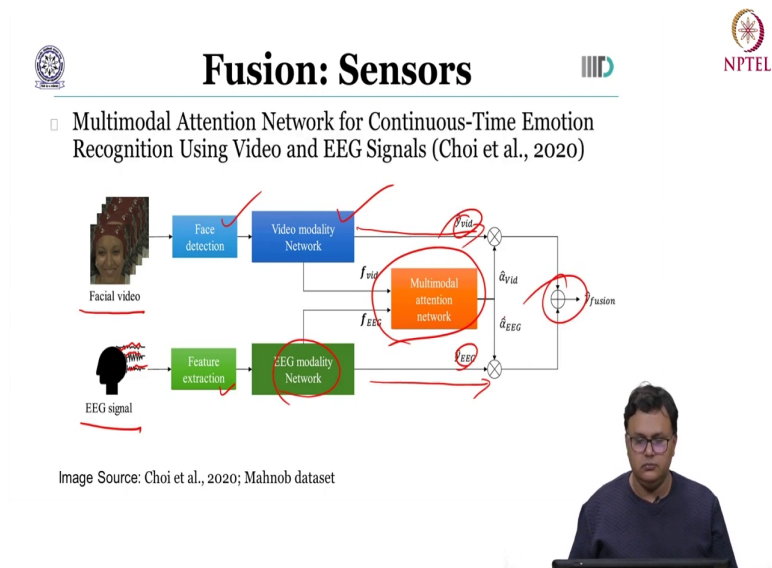
Image Source: Park et al., 2020



Now, this work is called ascertain and this is the very commonly used affect recognition and personality recognition dataset, when it comes to physiological data. Now, friends let us look at another work, where multiple sensors are used for affect. This is from Park and others. Here you see two subjects who are doing a conversation.

Now, the subjects have a head mounted camera, they have a mind wave headset, this is the EEG and then they have a heart rate sensor and the Empatica wristband. So, we are data coming in from different sensors. Now, this work called K-EmoCon. So, essentially this is a continuous emotion based data set, where the conversation between the two people, it happens in an spontaneous naturalistic fashion.




(Refer Slide Time: 49:59)




Now, if you look at one of the works for fusion with sensors, another work, here let us say you have the facial video, here you have the EEG signal, which is the object detection, we then have a pre-trained network.

In parallel, what we are doing is we extract the features which are coming from the EEG data, then we have a EEG network. What the authors propose is they are using attention network, where the input features from video and EEG are inputted. And further, the outputs you can see here, they are predicted for each EEG and video separately and then fusion is happening.

(Refer Slide Time: 50:51)





1. D'Mello, S., & Kory, J. (2012, October). Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies. In Proceedings of the 14th ACM international conference on Multimodal interaction (pp. 31-38).
2. Schuller, B., Wimmer, M., Arsic, D., Moosmayr, T., & Rigoll, G. (2008). Detection of security related affect and behaviour in passenger transport. In Ninth Annual Conference of the International Speech Communication Association.
3. <http://semaine.opendfki.de/wiki/SEMAINE-2.0>
4. Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., & Rigoll, G. (2013). LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. Image and Vision Computing, 31(2), 153-163.

Now, friends, this brings us to the end of the second lecture for multi-modal emotion recognition. What we have seen in this lecture is different methods, which have been proposed for affect prediction, when the data streams are coming from different sensors for both early and late fusion strategies.

Thank you.