

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

Week - 08
Lecture - 01
Multimodal Affect Recognition

Hello, I am Abhinav Dhall from Indian Institute of Technology, Ropar. Friends today, we will be discussing about Multimodal Affect Recognition. So, in the past few weeks, we have been discussing about single modality based emotion recognition systems. Today, we will see how we can combine them and get a more fairer, more accurate output with respect to the user emotion.

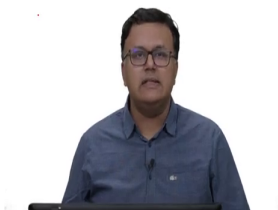
(Refer Slide Time: 00:53)



Multimodal Affect Recognition



- Why Multimodal?
- Types of fusion



So, in this lecture, first I will be discussing about why is multimodal presence of multiple modalities that is important and useful in a large number of circumstances. And then I will be discussing about the basics of fusion, which is how we are going to combine the information and their different methods and types.

(Refer Slide Time: 01:21)

The slide is titled "Motivation" and features several logos at the top: a circular institutional logo on the left, the IITD logo in the center, and the NPTEL logo on the right. The main content is a bulleted list with handwritten red annotations. The first bullet point states "Multiple modalities -> richer representation -> more accurate inference and expression." The second bullet point says "Each modality is expected to provide unique information," with "unique information" circled in red. Below it, "Redundancy" is circled in red. The third bullet point is "Challenges (D'Melo and Kory, 2012):" followed by three sub-points: "Classifiers and fusion methods that better capture the relationships between different modalities," "Affective corpora that contains adequate samples of synchronized expressions," and "Improvements are still often relatively modest." A handwritten diagram shows "Face (camera)" and "Speech (microphone)" connected by a double-headed arrow, with a red checkmark above it. Another diagram shows a vertical stack of three boxes with arrows pointing down to a larger box below, with "Camera Mic" written above it. In the bottom right corner, there is a small video inset showing the speaker, a man with glasses wearing a blue shirt.

Motivation

- Multiple modalities -> richer representation -> more accurate inference and expression.
- Each modality is expected to provide unique information,
 - Redundancy
- Challenges (D'Melo and Kory, 2012):
 - Classifiers and fusion methods that better capture the relationships between different modalities.
 - Affective corpora that contains adequate samples of synchronized expressions.
 - Improvements are still often relatively modest.

Now, the motivation is as follows. Let me try to enact. So, I am going to say I really like ice cream on a hot day. Let me say this again, but this time I am not going to face the camera, I really like ice cream when it is a hot day. And now, let me repeat this for the third time, I really like ice cream on a hot day.

Now, what has happened in this pursuit? In the first iteration, I was looking into the camera. So, you could analyze my facial expressions. In the second iteration, I was not looking into the camera, but you could hear me, right. So, the speech modality was present.

In the third iteration, as I was speaking, I was moving my head, looking down, looking sideways and what happened in this particular iteration? So, you had in some points, access to both voice and face and some points only the voice. So, that means, there was extra information, complementary information and sometimes vital information when one modality was present, the other was not present, right.

So, this gives us the motivation to actually have the possibility of using different modalities. And if you recall, the different modalities in this case friends is the camera based images and videos, the microphone based voice and music background and user information, the text based information in the form of messages, documents and then the physiological sensor based information coming from sensor such as the EEG or the ECG.

Now, this means we are going to combine multiple streams of data so that for our emotion recognition system, we can have richer information. Now, richer information would mean extra useful and complementary information which can help my machine learning model in better predicting the emotional state of the user.

Now, the accuracy is the concern here. From the perspective of when one modality could be present or could not be present as I showed to you and also when both are present perhaps, we could have a better performing system.

Now, the expectation of a multimodal emotion recognition system is that each modality is expected to provide some unique information. When you have these unique information from different modalities, then combining them makes a lot more sense.

Now, let us say in a scenario both the modalities face and voice are presenting the exact same type of information about the emotion. So, in that particular case maybe combining them

together would mean more computation, but not necessarily increase in the performance of the system.

Therefore, we create multimodal systems and during that process we keep this in mind that we are going to add those modalities to the system which would provide complementary information, unique information so that together the ensemble of this information coming from the different streams is useful for the prediction of emotion.

Now, we also do not want that if similar information is coming in then we end up with redundant data. So, along with compute there is a storage need as well. So, you end up requiring more storage and it is not really helping the system.

Now, according to Sidney D'Melo and Kory the challenges here for multimodal systems are as follows. Now, the classifiers the machine learning classifiers and the fusion methods the techniques with which we are going to combine the information coming from different sensors that better capture the relationship between different modalities. Now, what does that mean?

So, let us say you have face information coming from a camera and then you have the speech from the microphone you know you have the voice of the person let us say there is one user. Now, how do you combine them together such that we are able to better capture the information for emotion prediction and also what is the relationship between the data?

So, only once you understand the relationship between the data coming from different sources you can combine it better with the machine learning model or some statistical analysis method which you will see during the course of the lecture.

Now, from this challenge the second is the affective corpora. So, your datasets which contain adequate samples of synchronized expressions extremely important. Now, we would need a dataset for learning the patterns for emotion recognition. It is non trivial and time consuming process to set up an experiment for data recording where the face information.

Let us say here is the stream of face information coming in from the camera and synchronizing that with the voice information coming from a microphone, ok. So, they would have some time sync you know this was a timeline. So, this actually is a challenge to have large number of samples where we have the multiple modalities synchronized.

The other is it is observed in the large number of experiments that the improvement in the accuracy when you combine multiple modalities that might come out to be very modest, very minimal increase. Now, that could be based on the choice of fusion techniques, how much data do you have, how much complex machine learning or simple machine learning method you are using and also how many samples do you actually have for the multi modal data, right.

So, it is a challenge which comes into the picture that you know just because multiple modalities, face text information and you know so forth is available that does not mean that when you combine the different modality data the performance is going to go up you know this increase in performance of the system, increase in the accuracy of the system that is based on how the information is combined from different modalities.

And of course, as we already discussed this point is there any unique information which is there in the case of the different modalities. Because if it there is not unique information you combine the information together you may end up with the problem of curse of dimensionality from a machine learning perspective.

(Refer Slide Time: 09:55)



Multimodal Affect Recognition IIIID



- Underlying relationships and correlation between feature sets in different modalities and affect dimensions
 - Feature Selection
- How different affective expressions influence each other.
 - Uniqueness
- How much information each of them provides about the expressed affect.
 - Modality Selection

AFFECT
SHOGR

ACTIVE Feedback from user → Images ✓
Face + Body → Yes/No Binary answers



Now, let us look at multimodal affect recognition. What we are saying here is the underlying relationships and correlation between the feature sets from different modalities and affect dimensions this is what we want to understand, right.

Remember features here for example, for voice you could have your MFCC, for the face you could have a histogram of gradient and you would like to combine them, you would like to understand the correlation between them, study the relationship between them and that would help you in creating a multimodal affect recognition system.

Now, the first task here would be to choose the right features, feature selection, which feature should I use to represent the text? Which feature should I use to represent the physiological signals?

The other is how different affective expressions will influence each other? You know the presence of let us say face data and voice data or gesture data it will help one type of emotion category more to influence one category more than another category.

So, when you are designing a multimodal system combining the power of different representations from different streams you have to see you know which emotion category could gain more. And again now these are like a design choices which we are going to make when we are creating a system.

So, we are going to look for the uniqueness of the information with respect to emotion categories. Further how much information each of them provides about the expressed affect? Right, so let us say we want to create a system which is going to tell us that during the consumption of let us say massively online content you know some video content which is available on the internet, what is the emotion which is elicited in the user? Right, so you can now take two types of inputs.

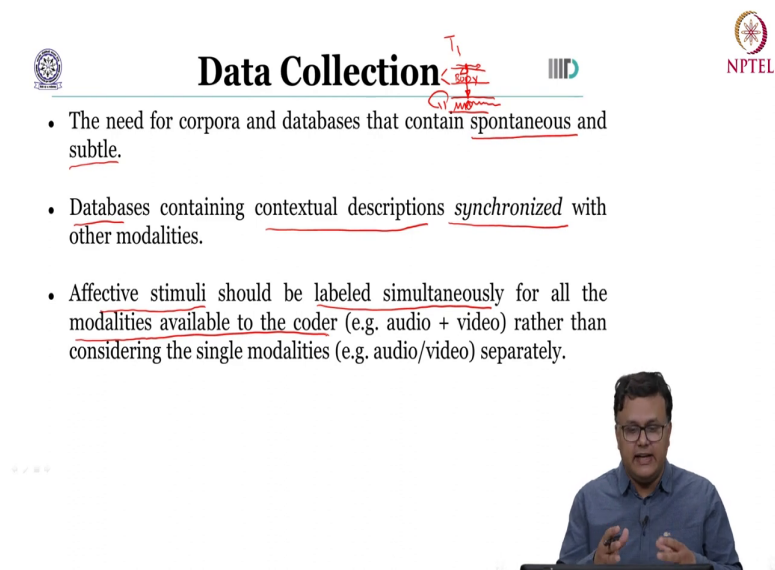
One let us say is you can capture the face and the body that is from the vision modality and this is active, right. So, you are. So, this is passive since you are recording this, ok. So, you are this is passive information. What you could also do is you could also have more active information wherein you can get the feedback from a user that after they watch the movie there were some questions and they answered, right. So, feedback from the user.

Now, these two are of very different nature the data, right. This is a continued stream of images and this one the feedback this could be simply let us say some yes, no binary answers.

Now, it needs to be seen that during the consumption, during watching of the video how much information can be provided by the images of the user. And later on let us say when you ask the questions from the user for example, did you enjoy watching this video, was this video informative? You get yes or no answers, right. So, this could also be indicative of the affect which was induced into the user when they watched the video, right.

So, during the consumption of the material the images are telling us better information, after the user has seen the video the feedback from the user is giving us more information, right. So, the multimodal system could combine this and can get a more accurate version of the emotion which was elicited into the user in this particular example. Now, you have selected the modality.

(Refer Slide Time: 14:21)



The slide features the title "Data Collection" in a large, bold font. To the left of the title is a circular logo, and to the right is the NPTEL logo. The title has handwritten red annotations: a "T1" above it, a "T2" to its right, and a "G" below it with a red arrow pointing to the word "Data". Below the title is a horizontal line, and under it are three bullet points. The first bullet point is "The need for corpora and databases that contain spontaneous and subtle." The second is "Databases containing contextual descriptions synchronized with other modalities." The third is "Affective stimuli should be labeled simultaneously for all the modalities available to the coder (e.g. audio + video) rather than considering the single modalities (e.g. audio/video) separately." In the bottom right corner, there is a video inset showing a man with glasses and a blue shirt speaking. To the left of the video inset, there are four small white dots.

Data Collection

- The need for corpora and databases that contain spontaneous and subtle.
- Databases containing contextual descriptions synchronized with other modalities.
- Affective stimuli should be labeled simultaneously for all the modalities available to the coder (e.g. audio + video) rather than considering the single modalities (e.g. audio/video) separately.

After you have selected the modality the question similar to how we talked about you know during the voice based emotion or the face based emotion prediction is the task of collecting the data. So, we will require let us say a labeled or a semi labeled dataset where the data would come from different modalities.

Now, of course, similar to unimodal when only single one modality is present, in the case of multimodal as well when we would be creating these datasets you know the expectation is

that they will have spontaneous data you know more natural data and it could be subtle as well.

Now, why is this important? Right, typically in the real world when we express ourselves most of the times the expressions are subtle. Now, these expressions could be in the form of the face information or the voice information only in very rare circumstances few cases you will see extreme expressions or you know extreme show of emotion in the voice or in the text.

So, the dataset which you would like to capture from different modalities that should contain spontaneous and should be representative of the real world circumstances. Now, what happens in the case of multimodal datasets which is different from your unimodal datasets your database it should contain this contextual description which is helping it through the synchronization with other modality, ok.

Now, what could that mean? Let us say a time stamp T_1 you have a stream of information coming; this is the image of the whole body of the user, ok. So, this is coming in and we are looking at this particular time stamp T_1 .

Now, in parallel at T_1 I also have the voice information coming in. Now, how do I know that exactly at T_1 which was the time stamp which could allow me to map these two together saying that ok, even though this is coming from a camera and this is coming from a microphone, how do I make sure that I get the information exactly at the same time for the different modalities?

Further during the data collection, I also need to take care of the stimuli. So, the stimuli it should be labeled simultaneously for all the modalities available to the coder; what that means is, let us say I am a labeler and I am going to now label assign emotion categories or emotion intensities if it was valence arousal scale to data which is let us say containing the camera modality and the voice modality. So, audio is there and video is there.

So, for the assigning of the emotional category as a label I have to be careful that I need to consider both the modalities the face and the voice, right. If I do individual level labeling then it will be non trivial to fuse them together. Because it is possible that if you only listen to the audio, it could convey a tad bit different emotion as compared to why you are just looking at the video, right.

So, while collecting multimodal databases we need to give very clear instruction to the coders, to the labelers about how they should interpret the data and what is the expectation on how the data would be interpreted.

(Refer Slide Time: 18:38)

The slide is titled "Feature Extraction" and includes the following content:

- Variable sampling frequency**
 - Video processing: 25 FPS | GSR: 16 Hz | EEG: 128 Hz etc.
- Synchronization is required**
 - To unite the feature information or
 - To come to a decision at a certain moment in time
- Feature Selection: Optimize the feature space individually per information stream followed by a combined feature selection** (Schuller et al., 2008).
 - I. Often highly correlated information should be reduced individually per modality

Handwritten annotations on the slide include:

- A diagram showing a 1-second interval divided into two parts, with 'V' and 'A' labels, and arrows pointing to 'F1' and 'F2'.
- A note: "discriminative info from the modalities unique info NORMALISE"

A video inset in the bottom right corner shows a man speaking.

Now, once we have created the dataset comes the next part which is extracting different features from different modalities. And now let us look at what are the important points which needs to be taken into consideration while we extract the features. Because in this case

features would be extracted from data streams coming from different type of sensors which means the data in different data streams would be of different nature, would be of different format, would be of different size. So, let us dwell into this.

So, what we will observe is, first thing there would be a variable sampling frequency. Now, let us say we had three sensors ok, the first is the camera, the second is your skin galvanic response you know that is let us say attached to the arm of the participant and then we have a headband. So, that is your EEG right, you put it on the head.

Now, if you notice typically the information coming from your camera would be of a frame rate let us say 25 or 30 frames per second, right. The information coming from your GSR and EEG that could be of a different frequencies. So, this one could be at a frequency of 16 hertz this fellow could be at 128.

Now, different strategies would be required to extract the features from these modalities individually. But since different amount of data is coming in a fixed unit of time across these different modalities, we will have to be careful with respect to how are we going to join the information together simply because we have different amount of information coming in.

Therefore, we are going to require methods for synchronization as well. So, let us say we say well, you know we are going to analyze 1 second of data together by analysis we mean we are going to extract the features.

So, I should know that from T_1 to T_2 which is of duration 1 second, how much of the video data has come in, how much of the GSR data has come in and how much of the EEG data has come in, ok. Because ultimately this will be the duration for which I am going to extract the individual features let us say F_1 , F_2 and F_3 .

So, I would require a way to synchronize. So, that I know that for the camera data which is coming in for the GSR sensor data and the EEG data what is this T_1 information, what is the

time stamp where in the stream of this data you have the starting point. And then let us say the ending point for this example where the duration we are taking is one second.


Now, after we would be synchronizing we will be able to combine these features, right. Let us say this is how we are combining. The other is we can also decide at certain moment in time that when do we want to make a decision, when do we want to have chunk of data analyzed coming from different modalities and then that would be let us say telling us the emotion category, right.

Further it is extremely important to do the right feature selection, right. It is about optimizing the feature space individually followed by a combined feature selection right, we have seen this already. So, when you have let us say the video coming in and then you have the EEG data let us say coming in then the optimum feature F 1 for this.



The optimum feature for EEG, then an optimum strategy for combining this F 1 and so that F 1 and F 2 give you the most discriminative information about the current state of the user and this should have the unique information from the modalities in this case you know we have to.

And if they are unique, we also need to normalize them as well. Because the video modality for example, F 1 and then the feature from video modality the EEG modality F 2 they would let us say be of different ranges right, different ranges, different sampling rate. So, we have to normalize them and have the combined feature in the most optimum manner.

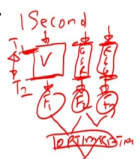
(Refer Slide Time: 23:42)



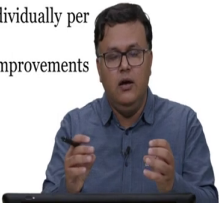
Feature Extraction



- Variable sampling frequency
 - Video processing: 25 FPS | GSR: 16 Hz | EEG: 128 Hz etc.
- Synchronization is required
 - To unite the feature information or
 - To come to a decision at a certain moment in time
- Feature Selection: Optimize the feature space individually per information stream followed by a combined feature selection (Schuller et al., 2008).
 - I. Often highly correlated information should be reduced individually per modality
 - II. Secondary optimization process can lead to further improvements reducing cross-modal redundancy



1 Second

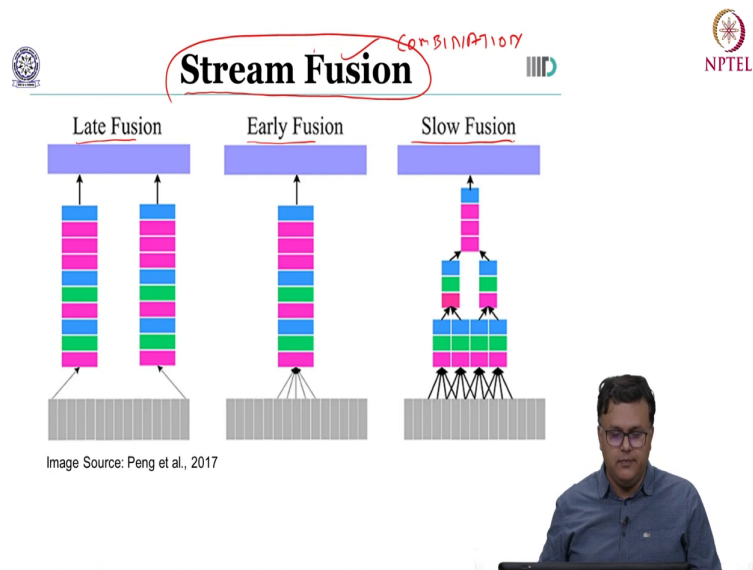


It is often noted that highly correlated information would be reduced individually per modality, right. So, since we are looking at the unique and of the correlation, we would be highly dependent on the individual modality as well, the quality of individual modality and the quality of feature which is representing that individual modality data.

Now, further once you have combined let us say the features right, you combine these two features you could do certain optimization at this level, right. So, one optimization of course, is the normalization, it could also be let us say some dimensionality reduction right, or projecting the data on to a higher dimension space such as simply we reduce the cross modal information.

So, we do not want the information which is same in phase and in EEG or in EEG and in speech, what we want is the complementary information. So, that together its more information not just you know more data which has redundant information.

(Refer Slide Time: 24:55)



Now, once you have selected the type of feature representation, the next is how do we combine them together? And combining is here called the fusion process. Now, broadly friends, we can divide the fusion process into late fusion early fusion and slow fusion. Now, we will study each one separately.

(Refer Slide Time: 25:33)

Early Fusion

- Concatenating the the features from multiple cues into one feature vector.
- Becomes more challenging as the number of features increases and when the features are of very different nature.
- Synchronization is of utmost importance.
- Less information at fusion level?

The diagram illustrates the early fusion process. At the bottom, three separate feature vectors are shown: $[F_1, F_2, F_3]$ (labeled 'FACE FEATURE'), $[T_1, T_2, T_3]$ (labeled 'TEXT FEATURE'), and $[S_1, S_2, S_3]$ (labeled 'SPEECH FEATURE'). These are combined into a single 'COMBINE FUSE' vector. This fused vector is then processed by an 'ML PROCESSING' block, which outputs a final feature vector. Handwritten notes include 'NORMALISE' and 'SAME DIMENSIONED FEATURE'.

Image Source: Peng et al., 2017

Let us start with the early fusion the simplest fusion method. So, what we are going to do is, we are going to concatenate the features from multiple modalities into a single feature vector. Now, if you see here let us say what is happening is, let us say we have face feature, then we have some text feature and some speech feature, ok. So, there are three modalities

What I am going to do is I am going to take these three features and combine them, fuse them, ok. How do I fuse them? The simplest approaches you append let us say feature 1, and then you behind feature 1 you put feature 2, then append a feature 3, right. And then here you have some machine learning processing could be or any machine learning algorithm here we are showing a neural network in the illustration, but you could have any other machine learning method.

Now, in this case it becomes challenging as the number of feature increases, right. And also, when the features are of very different nature; quite obvious. What we are doing is we are combining the features, right. When we are combine the features, if let us say we had multiple features coming in from different modalities. Well, we will have a large feature and also if the features are very different, now what does that mean?

Let us say F 1 face feature is for one frame and is of dimension 1 into 768 you know this is just an hypothetical example, F 2 is a text feature which is actually of the dimension it is a matrix, ok. So, let us say it is a 5 cross 12 and then speech feature is actually a spectrogram, ok. So, again you could say well that is let us say of the dimension 128 cross 128, this one is one dimensional, this one actually is matrices, right.

So, when the features are very different how are we going to fuse? One could say well, you know I am going to fuse it this way F 1 then I take you know flattened F 2 and then I take flattened F 3 and I simply append them. But this might not be the most appropriate, the most useful type of fusion. Because not only we would end up with a very long large feature, but it is also possible that the values are quite different coming from different ranges in this, right. So, you will had to normalize as well.

Now, in the case of early fusion friends synchronization would be extremely important, right. Because we are combining the information at the feature level itself. So, as I discussed earlier, we will have to know the starting point and the ending point in time and we will have to take the information exactly from those starting and ending points across the information coming from different modalities.

Now, one could say well, one way to combine the features coming from different modalities is simply concatenating them. There are other interesting yet simple ways one could not say well, if for example, the dimension of the features coming from different modalities that is similar post normalization one could extract simple statistics. Let us say I convert the different features coming from different modalities into same dimension feature. So, I do some pre processing.

(Refer Slide Time: 29:52)

Early Fusion

- Concatenating the the features from multiple cues into one feature vector.
- Becomes more challenging as the number of features increases and when the features are of very different nature.
- Synchronization is of utmost importance.
- Less information at fusion level?

(F1, F2, F3) NORMALISE

SOME DIMENSIONED FEATURE

MEAN STD

ML PROCESSING

COMBINE FUSE

FACE FEATURE TEST FEATURE




Image Source: Peng et al., 2017

The diagram illustrates the early fusion process. At the bottom, there are two boxes labeled 'FACE FEATURE' and 'TEST FEATURE'. Arrows from these boxes point to a central box labeled 'COMBINE FUSE'. Above this box, there is a vertical stack of colored bars (blue, green, yellow, red, purple) representing the combined feature vector. An arrow labeled 'ML PROCESSING' points from this stack to a box at the top labeled 'Emotion'. The slide also includes handwritten notes: '(F1, F2, F3) NORMALISE' and 'SOME DIMENSIONED FEATURE' near the input features; 'MEAN STD' near the 'COMBINE FUSE' box; and 'ML PROCESSING' near the 'Emotion' box. The image source is cited as 'Image Source: Peng et al., 2017'.

Then what I can do is since now the different features are of the same dimension coming from different modalities, I can compute things such as the mean and the standard deviation, right. So, this could become my feature which goes in to the machine learning model for the further processing for getting the emotion level.

Now, another challenge and a limitation which comes in for your early feature fusion for multimodal emotion recognition is that since we are combining the features together, we are actually not analyzing the features for their discriminativeness or unique information first and are simply doing a combination, right.

(Refer Slide Time: 31:04)

 **Late Fusion**  

- Feature and time dependency are abstracted.
- Each classifier process its own data stream and the multiple sets of outputs are combined at a later stage to produce the final results.
- Soft level: a measure of confidence is associated with the decision
- Hard level: the combining mechanism operates on single hypothesis decisions.

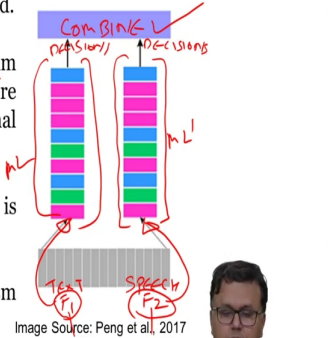


Image Source: Peng et al., 2017

So, we may lose a lot of temporal information in this pursuit, right. The individual transitions which are happening between the data, coming from individual modalities that might be lost.

So, now let us look at the late fusion friends. In the case of late fusion as the name suggest what we are doing is we say well, let us say we have now two features coming in, ok. One feature is your text data and the other is your speech data, ok. So, you have feature F 1 and you have feature F 2.

What you are going to do is you are going to take F 1 and you are going to input that into your ML pipeline, ok. You separately take F 2 and you input that into another pipeline. Now these can be similar or these can be different depending on the nature of the feature and the complexity of the feature and what type of information you want to extract.




Now, what you are going to do is you are going to get some decisions, ok. Let us say from your ML pipeline one you get some decision about the emotion and from the second pipeline ML dash you also get some emotion decision.

Now, notice this is for one data point you have the text and speech, you input that individually separately into the machine learning pipeline and you got some emotion decisions. Now you are going to combine these decisions, ok. So, the combination the fusion happens late into the pipeline of processing and that is why we call these type of systems as late fusion based multimodal emotion recognition. So, each classifier processes its own data stream, the multiple set of outputs are combined as we did here at a later stage.

Now, in this case there are further two sub-categories. The first is we do late fusion at a softer level which essentially is simply saying well you know we use a measure of confidence which is associated with the decisions. The second is late fusion at hard level which is simply saying combining the mechanism and operating on a single hypothesis decision.

So, in one case you are taking the probabilities let us say which are coming from the two different channels as in here in the example. And in other case you are simply combining the decisions. A very simple example is let us say you take the decisions about the emotions and you do AND or a OR across these two and that is going to give you the final emotion, ok.

(Refer Slide Time: 33:46)

 **Slow Fusion**  

- Assumption of conditional independence between modalities and cues in decision level fusion can result in loss of information.
- Exploit the correlations between the modalities while relaxing the requirement of the synchronization.

CapL!

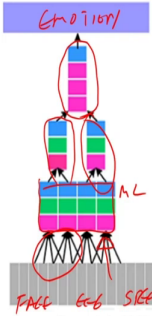



Image Source: Peng et al., 2017



Now, related to both early and late fusion in a way you know a middle ground between early and late fusion is friends your slow fusion process. What you are saying here is let us say again you know you have face and you have your EEG information or maybe you know some speech information coming in.

You are going to combine the information multiple levels, ok. Perhaps let us say you know you combine face and EEG together here speech goes in individually and you have a common ML system, ok. Further what you are doing is you are taking some part of the system and you are individually analyzing it, let us say you get the decisions, you again learn a small classifier on the decisions and that gives you the final emotion.

So, what is happening? There is an assumption of conditional independence between the modalities and in the cues of decision level fusion. What happens when you are actually

doing just decision level fusion you know the late fusion we may have already lost a large amount of information which could be gained by let us say computing the correlation or understanding the correlation between the individual features.

Because when you were doing late fusion you got the decisions for all the modalities separately and then you combined them, you did not analyze the features together, but you analyze the decisions together, right. But here we have the best of both the worlds, we are actually not only harbouring on to the independence the uniqueness of the features, we are also computing the correlation between the features and then we are also combining the decisions as well.

So, late slow fusion gives us the correlation between the modalities while relaxing the requirement of synchronization. So, as compared to your early fusion we are synchronization is extremely important, in this case you know we can relax a bit because we are combining at different stages.

So, friends with this we reach towards the end of lecture 1 of the multimodal affect recognition systems. We looked at why multimodal systems are useful, what are the challenges in creating multimodal systems and later on we discussed about the different strategies in creating the multimodal fusion system with respect to the early late or slow fusion.

Thank you.