

Affective Computing
Dr. Abhinav Dhall
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi

Week - 04
Lecture - 01
Automatic Facial Expression Recognition

Welcome to the lecture on Automatic Facial Expression Recognition as part of the affective computing course. Now, imagine you are talking to a friend. The friend tells you about how his or her they went. Now, just by looking at their facial expression, their body gesture and listening to them through the speech, you can tell what is their mood.

Did they have a good day? Are they comfortable in the conversation? And based on that, you are going to give an apt reply. Now, we want to have the same in a machine. For that, we actually look at the methods of automatic facial expression recognition. Now, what will that mean?

We want to understand the emotion of a user or a group of users. For that, we will be looking at the cues which the user is giving to the machine. Now, these cues again will be in the form of their facial expressions, their head gestures as you can see me now, I am right now nodding and also their body gestures.

So, how can we track this information and then tell a user interface that how is the user reacting to the system? And in this process, try to have a more efficient interaction with the system. Now, emotions are very deep embedded in our minds. Whenever we are trying to interact with someone, the cues which we get from them give us some idea about their emotion.

However, without the context, it is difficult to understand someone's emotion. But the nearest neighbor to that is the facial expression. Right. One can say that during conversation, when you are looking at a facial expression of a person that can tell you give you a glimpse of the

emotion of that person. Therefore, we are going to do two tasks in automatic facial expression recognition.

(Refer Slide Time: 02:23)



The slide features the NPTEL logo in the top right corner and a university logo on the left. The main title is "Facial Expression Recognition" with a blue bar underneath. Below the title is a bulleted list: "Automatic Facial Expression Recognition (FER)", "Emotion Recognition", and "Expression Recognition". To the right of the list is a photograph of a woman's face with red annotations: a red box around the face, red lines for the upper and lower lips, and a red circle around the word "HAPPY" in the top right. A vertical stack of labels "AUG", "AUG", "AUG", "AUG" is on the right side of the face. In the bottom right corner, there is a small video inset of a man in a blue shirt speaking.


The first is your emotion recognition. We want to understand the deep emotion of the person. The second is your expression recognition. Assuming that the facial expression of a person is the window to their emotion, we want to understand and then categorize the different facial expressions which can tell a machine how is the user feeling.

Now, friends, an example of that is on the slide. You see a person here. So, the face is detected. And based on the lower lip, the upper lip. And if you notice the movement here around the chins and the eyes, one can tell that this person is happy. Now, if this person is happy, the system can react accordingly.


One can also go a further level down and then look at what are those muscles which are activated. Now, these are action units which correspond to the different facial parts. And in the later part of my lectures, I will be discussing with you about facial action units as well.

Now, typically to understand emotions, there are a wide variety of sensors which are available. Some of them are obtrusive, that is you will put a sensor on the person. Some of them are non- obtrusive. In today's class, we will be focusing mainly on the non- obtrusive sensors.

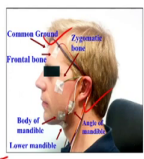

(Refer Slide Time: 04:00)



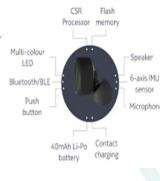
Facial Expression Recognition |




- Automatic FER
 - Emotion Recognition
 - Expression Recognition
- Inputs to FER:
 - Electromyograph (EMG)
 - Electrocardiogram (ECG)
 - Electroencephalograph (EEG)
 - Earables
 - Camera
 - Most informative
 - Non-intrusive

wearables





Now, the first one is a EMG, your electromyograph. Then you have your electrocardiogram which is your ECG. And then you have the electro encephalogram which is the EEG. So, EMG, ECG and EEG as you can see here are actually put on the body of the user.

Now, what; that means, is the user could be aware that there is a sensor which is attached to the body of the user. Therefore, if he or she is aware that the system is trying to understand their emotion, they may be giving bias signals. Of course, when you look at these sensors, they have been improved.

You already have the variables these days. For example, the smart watches, they become so, ubiquitous with the user that the user does not really always actively analyze and realize that they are trying to understand the physiological meanings of the body of the user. For example, the heart rate and the blood pressure right.

What we are more interested today is the camera based signals. Now, you attach a camera to let us say the wall in a room and then you would like to understand the expressions of the person. The usefulness is as the facial expressions are the gateway to the emotions of a person.

You can get a large amount of data at a very low cost. Camera sensors are not very expensive these days. You get cameras in different forms, from your CCTV cameras to the camera phones in the smartphones in the front and back. Further, the biggest advantage is the non-intrusive nature.

You can have a person being analyzed by putting a camera which is further away from the person. Of course, that does not mean that the obtrusiveness does not harbor into their personal space, but for the sake of discussion, let us say its very non obtrusive.

(Refer Slide Time: 06:26)



Static and Dynamic FER



- Static facial features obtained by extracting handcrafted features from selected peak expression frames of image sequences.
- Spatio-temporal features to capture the expression dynamics in facial expression sequences



Now, when we are further going down into facial expression recognition, there are two important categorizations which I would like to make. The first is based on the amount, the type of data which you are receiving, you would like to categorize a particular face into a subset of expressions.

The first one is your static facial expression recognition. Now, as the name suggests, you capture one frame containing a person and you analyze the information in that frame to predict the expression and later the emotion of the person. And whenever we are saying emotion, please notice most of the time we are referring to the perceived emotion.

Perceived emotion is you are talking to someone, now what is your understanding of the emotional state of that person? Of course, the other side of that is that a person tells you yourself that this is how I am feeling, so, that would be the self-labeled emotion. Now,

typically what you will do in static facial expression recognition is, you have a frame and then you would extract some statistics, some features from that.

Now, obviously, the features would correspond to the different shape which you see in your face. For example, if you notice me right now, I am smiling. Now, as I smile, the lips, the muscles around that they elongate, right. So, we would need a feature, a representation which can tell the difference between a neutral versus a smile.

Now, typically these frames for static facial expression recognition are selected from a series of frames which are captured by a camera. One may also say that I will use the selected peak frames, ok there are different methods with which you can you choose a frame.

For example, one could say, well, I would like to select one frame every second and then analyze the expression. Another way could be that we would like to only choose a frame to analyze the expression when there is a substantial movement or substantial change which has happened in a frame.

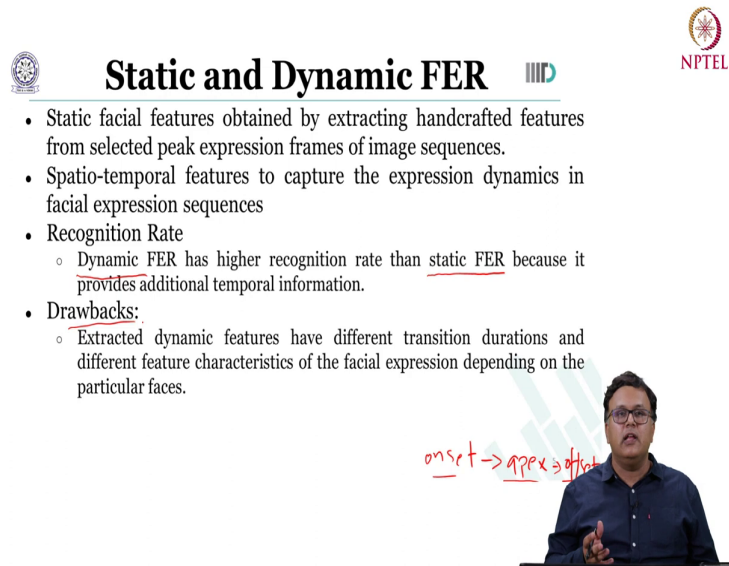
Now, this idea of course, is borrowed from the world of video compression, where we are looking for the change in terms of the frame and we will use one frame as a reference frame. So, the peak expression frame is actually a reference frame. Now, let us say you have a series of frames coming in, why not actually analyze the series of frames together and do what we refer to as spatiotemporal facial expressions.

Since expressions are dynamic in nature, what does that mean? Now, again if you notice me friends, I am going to smile. Right so, when I started to smile, there was an onset of the smile and then the expression reached a peak and then there was an offset. So, there is a temporal movement which is happening through the expression.

Since in the context of frames, you have movement happening in the lip region and the other parts of the face in both space that is at the frame level. Let us say this is the face at the frame

level and at the other frames level as well. So, spatio this is t, this is your smiley spatiotemporal ok.

(Refer Slide Time: 10:36)



The slide features a title "Static and Dynamic FER" with a small logo to its right. To the left of the title is a circular institutional logo. To the right of the title is the NPTEL logo. Below the title is a bulleted list of points. A video inset in the bottom right shows a man speaking, with handwritten red text "onset -> apex -> offset" overlaid on the video.

Static and Dynamic FER

- Static facial features obtained by extracting handcrafted features from selected peak expression frames of image sequences.
- Spatio-temporal features to capture the expression dynamics in facial expression sequences
- Recognition Rate
 - Dynamic FER has higher recognition rate than static FER because it provides additional temporal information.
- Drawbacks:
 - Extracted dynamic features have different transition durations and different feature characteristics of the facial expression depending on the particular faces.

Now, further it has been observed in a lot of studies that this spatiotemporal facial expression which is also referred to as your dynamic facial expression recognition, it achieves a better recognition performance as compared to a single frame based static facial expression recognition.

The reason is quite obvious expressions are dynamic and hence, when you are analyzing a series of frames, you get extra information about the onset, apex and offset of the frame and these dates of the expression will then give you information about not just what is the expression, but also about when did the person started, let us say to smile and when did that smile episode end.

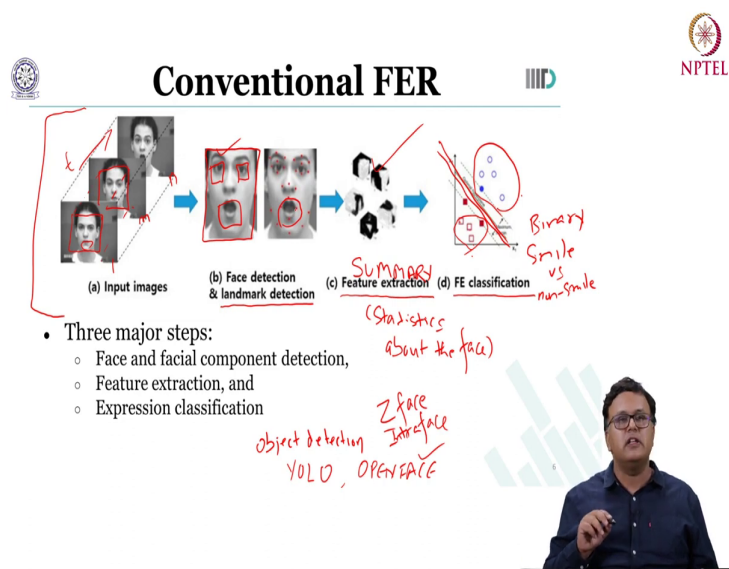
It is extremely important to be able to localize this expression as well because one thing is to understand the expression and the perceived emotion of a person and the other thing is once you understand the expression and the emotion, how you want your machine to react? Now, I have already told you the advantage of dynamic facial expression, the spatiotemporal analysis as compared to static facial expressions, there is a drawback.

When we are extracting dynamic features from series of frames, they can have different transition durations and different feature characteristics depending on particular faces. Now, different people will have a different reaction to the same stimuli, same stimuli means, let us say a joke is cracked in front of two people, one person can laugh more than the other person.

Now, what; that means, from an affective computing perspective is, the laugh event will have different durations for the two persons. If they are going to have different durations, that means, your system which is extracting some type of features, some statistics around that event needs to be agnostic to the different durations, small laugh, long laugh.

Further, of course, when you are analyzing the information from the perspective of series of frames, you are expecting a lot more compute requirement. As when you are combining these frames, there will be a lot more data, simply means we are going to need to have a more powerful machine.

(Refer Slide Time: 13:33)



Now, from the context of both static and dynamic facial expression recognition, here is a typical pipeline. In the beginning, here you see we have a series of frames. So, this is t and what you are doing is you are capturing these frames and later, you want to detect the location of the face.

So, where is the location of the face? Now, this location of the face can be detected using various object detection techniques. So, if you look at open source systems, there are methods such as YOLO, which is you look only once. Then there are face specific object detection systems, for example, open face. So, that is an open source face detector and face tracker as well.

So, what does that mean? You detect the location of the face in the first frame and then in the consecutive frames, m th frame and n th frame, you are tracking the where has this object

moved across corresponding frames. So, that is a tracker. Now, once you have the face localized, that is the location of the face in an image, you would be interested in understanding the location of the different facial paths, the eyes, the lips and so, forth. For that, we will do landmark detection.

Now, again, for landmark detection, there are a large set of open source libraries. For example, open face does that, there is one called intraface and then there is another one called Z face. So, you can detect and then you can find the landmarks, the location of the different facial points in my image.



Now, once we know where the facial paths are, we would be interested in extracting features. Now, these features are also some statics statistics about the face, which essentially are going to create a summary. Now, this summary is going to tell me information about let us say if this is open or in another frame, for example, here the mouth is closed.


Now, once you have extracted this summary, essentially you have the feature, you are going to use a machine learning algorithm to do classification. Now, here what you are seeing essentially is that let us say this is all the faces, all the data points representing one face. So, this is one face, this is another, this is another and let us say all other points which represent face where the person is not smiling.

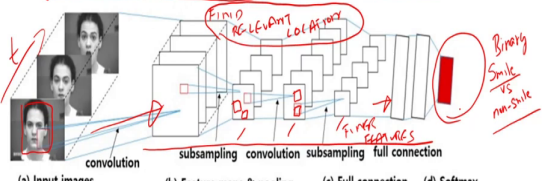
So, you can say well, I have a binary task which is smile versus non-smile. So, I would be learning this boundary which can segregate the smiling faces from the non-smiling faces. Now, of course, there are a large number of factors which are going to go into the optimal boundary.

From the context of affective computing, what will that mean is, you need to have the right facial expression feature and we will be studying some of the features in the coming slides. The other is you need high quality localization of the face and the facial points. So, if you were to extract some statistics around the lip region, you would need good localization. That means, precise location of where the facial points are with respect to the lip region.

(Refer Slide Time: 18:12)


 **Deep-learning-based FER** 





(a) Input images (b) Feature maps & pooling (c) Full connection (d) Softmax

- Deep-learning-based FER approaches:
 - highly reduce the dependence on face-physics-based models and other pre-processing techniques
 - Enabling end-to-end learning to occur in the pipeline directly from the input images



Now, this was when you are doing conventional facial expression recognition. From 2012, we have seen tremendous progress in machine learning and related areas due to the deep learning based approaches. So, of course, you can do automatic facial expression recognition with neural networks. Now, what will that mean? Again, here you have a series of frames.

And what you are doing is, you are taking these frames and you are giving it as an input to a neural network. Now, this neural network, it is going to learn these feature representations which are going to correspond to the final goal here. Now, let us say again, similar to our earlier example, we are doing smile versus non-smile. So, for this binary task, you have a set of these samples and then you learn a convolutional neural network. Ok.

Now, it is going to learn these feature based on the filter weights. So, these filter weights are going to be learned based on the trained data. And what; that means, as compared to the

earlier hand engineered base feature approaches is. So, the system will not only learn the larger shape changes, but subtle expressions as well which can help it in better doing classification. Smile verses non-smile.

Now, this is also referred to as sometimes an end to end learning approach. What; that means, is, you input the image directly. So, no longer you are doing face detection. In some cases, you input the image directly because you know that most of the component of my image already is a face.

You can expect then the network to learn for example, here to find relevant location. So, it will find relevant location in the image and the early on layers will then give more importance to the relevant location which is the face. And then the later layers will extract finer features.

This finer features are the location let us say of the lips, the location of the eyes. And that will further go into your fully connected layers and later you will have your classification. Now, this deep learning based facial expression recognition is very prevalent. It has shown to perform much better as compared to the traditional hand engineered based features.

Of course, there is a subtle drawback here. The drawback essentially is that as with most of the supervised neural network approaches, you would need a large amount of training data. Now, collecting trained data is also a non-trivial task when it comes to facial expression recognition and we will discuss some aspects of that later.

The other of course, is the high energy requirement which you would need for not just the training of your deep neural network, but also the inference time requirement. So, in the end you will end up using more energy as compared to your traditional hand engineered features.

So, what; that means, is there is a trade-off between your hand engineered facial expression recognition based systems and deep learning based facial expression recognition systems and the trade-off is simply based on the accuracy which I need versus the energy which I am ready to spend for achieving that particular facial expression recognition accuracy.

(Refer Slide Time: 22:55)

UNIVERSAL EXPRESSIONS

FROWN
DISGUST
FOR
HAPPY
NEUTRAL
SAD
SURPRISE



Macro Expressions

- Macro Expressions
 - Used primarily to capture obvious/universal facial expressions.
 - Visually observed through facial landmarks which are salient points in facial regions such as the end of the nose, ends of the eye brows, and the mouth etc.
 - Last between 1/2 a second to 4 seconds
 - Match the content and tone of what is said



Now, once we have the classification of facial expression recognition systems into your static versus dynamic facial expression recognition and later on hand engineered versus deep learning based facial expression recognition system. Let us talk about how are we going to represent the expressions.

So, the first one here are your macro expressions. What are the macro expressions? Now, these are the obvious easily understood expressions. You talk to a person, you see them smiling and you say well the person is happy because he or she is smiling. Now, you look at these image simply these are the obvious expression this person is happy.

This person is you know sad and here you actually have more of a neutral type expression, a bit of you know contempt as well. Now, these macro expressions these are visually observed

through the major facial locations. Now, these are also referred to as the salient points in a phase. What are the salient points?

Now, these are the points which are giving us the movement of the different facial paths. Also, sometimes referred to as the non-rigid part of the phase. These typically would last through half a second to a 4 second duration and given that their macro expressions they will match the content and the tone of what is being said.

An example is as follows. I am very happy to speak to all of you for about facial expression recognition. Now, when you look at my expression and you hear my tone you can correlate that right. Now, further there has been a lot of study about what these macro expressions can be.

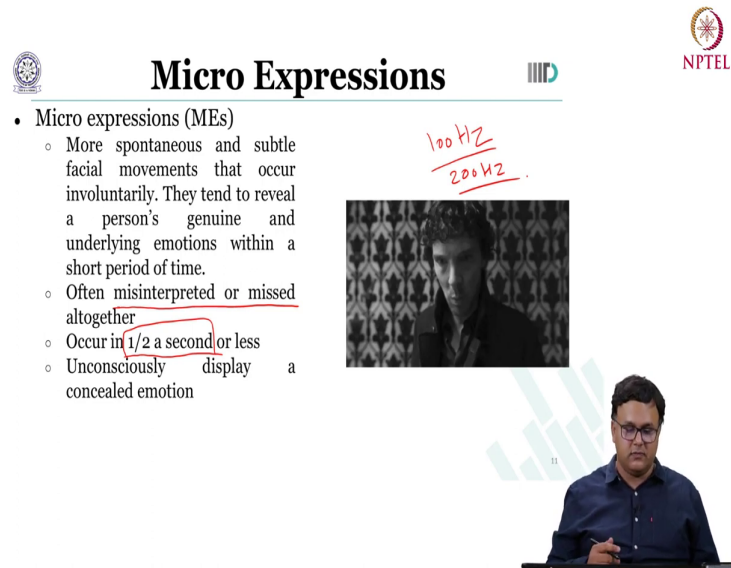
The first one which is a bit simpler representation of this macro expression is a universal expressions. Now, these typically would be your anger discussed, fear, happy, neutral, sad and surprise. For a couple of decades, it has been assumed that these are the universal expressions you know Irrespective of the nationality, irrespective of the culture from which a person comes, these are always observed against the same type of stimuli.

Only recently it has been found that not all these expressions, the 6 plus neutral universal expressions are found across all the cultures. Anger, happy, surprise, these are commonly found across cultures. But you would have noticed that people coming from different regions of the country or from different cultures will represent their emotions in different ways.

So, these are no longer universal, but still an easy way to represent the facial expressions. And of course, as with any affective computing problem, these are based on the ultimate understanding which you want to have in terms of the affect, emotion of the person. So, depending on how serious the use cases, we can choose a subset or all of these universal expressions.

And as the name suggests for this universal macro expressions, it is easier to collect data because we are only creating one label for an image. For example, this image has the subject showing happy expression.

(Refer Slide Time: 27:28)



The slide features the title "Micro Expressions" in a large, bold font. To the left of the title is a small circular logo, and to the right is the NPTEL logo. Below the title, a bulleted list describes micro expressions (MEs). The list includes: "More spontaneous and subtle facial movements that occur involuntarily. They tend to reveal a person's genuine and underlying emotions within a short period of time.", "Often misinterpreted or missed altogether", "Occur in 1/2 a second or less", and "Unconsciously display a concealed emotion". A red box highlights the phrase "1/2 a second or less". To the right of the text is a small video frame showing a man's face with a subtle expression. Handwritten red text above the frame reads "100 Hz" and "200 Hz". Below the video frame is a small inset image of a man in a dark shirt and glasses, likely the presenter.

- Micro expressions (MEs)
 - More spontaneous and subtle facial movements that occur involuntarily. They tend to reveal a person's genuine and underlying emotions within a short period of time.
 - Often misinterpreted or missed altogether
 - Occur in $1/2$ a second or less
 - Unconsciously display a concealed emotion

Now, if you want to develop a bit deeper, there is a concept of micro expressions as well. Now, what are micro expressions? Micro expressions are spontaneous and subtle facial movements that occur involuntarily. Please note, these are the ones which occur involuntarily. And they tend to reveal a person's genuine and underlying emotion within a very short period of time.

So, as compared to macro expressions, your micro expressions, they are closer to the emotion. Now, what are your micro expressions? For example, you notice this image here.

There is a subtle twitch here. There is a movement of the eye here. Now, these are the ones, for example, let us take a use case.

There is an interrogation happening. Person 1 asks person 2, a difficult question. Person 2 wants to hide the information. Their reaction right away after the question has been asked will be a combination of some involuntary muscle movement and some voluntary muscle movement.


Now, this involuntary short duration muscle movement is your micro expression ok. You would have seen this in popular science TVs as well. Right the investigators are able to tell the you know the person was telling truth or was trying to hide something. Now, for that they are trying to analyze the micro expressions.

As the name suggest, it is non-trivial to understand to interpret micro expressions. The reason for that is they have a very short duration. Typically, they are less than 500 milliseconds. Now, what; that means, is to a naked eye, to an untrained expression expert, it is difficult. Even for an expert who has been trained to understand micro expressions, they will really have to focus because of the duration.

What that also means, is when you are going to create a dataset for micro expressions, it is going to be more difficult to create and then label these expressions. So, typically what will be done is the cameras which are used to create these large datasets where micro expressions are labeled, they are of very high frame rates.

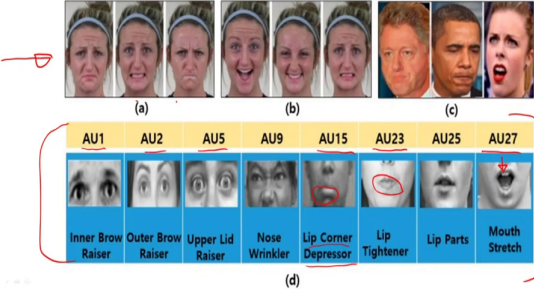
For example, you will see some datasets coming from the University of Oulu in Finland where they will have 100 up to 200 hertz. So, you are actually you know having a lot more data a because the event is small, you record more frames and that way you have sufficient amount of information which later you can use to extract your features.









(Refer Slide Time: 30:40)




Facial Action Coding System

- Facial action coding system (FACS) DU1 → DU27 →
 - A system based on facial muscle changes and can characterize facial actions to express individual human emotions (Ekman and Friesen, 1978).



AU1	AU2	AU5	AU9	AU15	AU23	AU25	AU27
							
Inner Brow Raiser	Outer Brow Raiser	Upper Lid Raiser	Nose Wrinkler	Lip Corner Depressor	Lip Tightener	Lip Parts	Mouth Stretch

(d)



Now, the third one friends is your facial action coding system and also referred to popularly as the FACS system. Now, after Ekman proposed the universal expressions, he and Friesen also proposed the facial action coding systems. Now, what are facial action coding systems?


Well, these are the facial muscles which would get activated when a person shows a particular expression. So, typically let us see your smile will be combination of different facial muscles ok. Now, here you see a set of frames. Now, these different expressions are essentially combination of different facial muscles moving in which means different facial actions have happened.

Now, here you see a subset of facial expressions. So, facial expression based action unit 1, AU2, AU5, to AU27 and their correspondence as well. And notice how localized they are. They are not talking about the whole face, they are only talking about a local region. So, for


example, when you see these lower lip movement, you have AU15 which corresponds to lip corner depressor.

When you have AU23, that would mean that the user has you know tightened their lip. Ok. Now, what we will typically do is, you know we will actually let us say combine certain AUs. So, maybe AU1 plus you know AU27 and then you could say this corresponds to a certain expression.

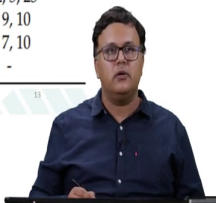
(Refer Slide Time: 32:31)



AUs (Benitez-Quiroz et al., 2017)



Category	AUs	Category	AUs
Happy	12, 25	Sadly disgusted	4, 10
Sad	4, 15	Fearfully angry	4, 20, 25
Fearful	1, 4, 20, 25	Fearfully surprised	1, 2, 5, 20, 25
Angry	4, 7, 24	Fearfully disgusted	1, 4, 10, 20, 25
Surprised	1, 2, 25, 26	Angrily surprised	4, 25, 26
Disgusted	9, 10, 17	Disgusted surprised	1, 2, 5, 10
Happily sad	4, 6, 12, 25	Happily fearful	1, 2, 12, 25, 26
Happily surprised	1, 2, 12, 25	Angrily disgusted	4, 10, 17
Happily disgusted	10, 12, 25	Awed	1, 2, 5, 25
Sadly fearful	1, 4, 15, 25	Appalled	4, 9, 10
Sadly angry	4, 7, 15	Hatred	4, 7, 10
Sadly surprised	1, 4, 25, 26		-



Here are some of the mapping friends. So, let us say when you have the facial expression which is happy, then action units 12 and action unit 25, they are activated. Let us look at the other side, negative expressions. So, when the category type is sadly disgusted, you have the action units 4 and 10 with are activated.

So, ultimately from a facial expression recognition system perspective, that will mean you need to first detect the action units. And then you can use this information to go deeper in things such as perceived emotion or you can use action units to let us say predict things such as observe pain in a patient. So, you can train a system for that. Now, we will stop here and in the next lecture, we will be covering the aspects of facial expression recognition.