Affective Computing

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

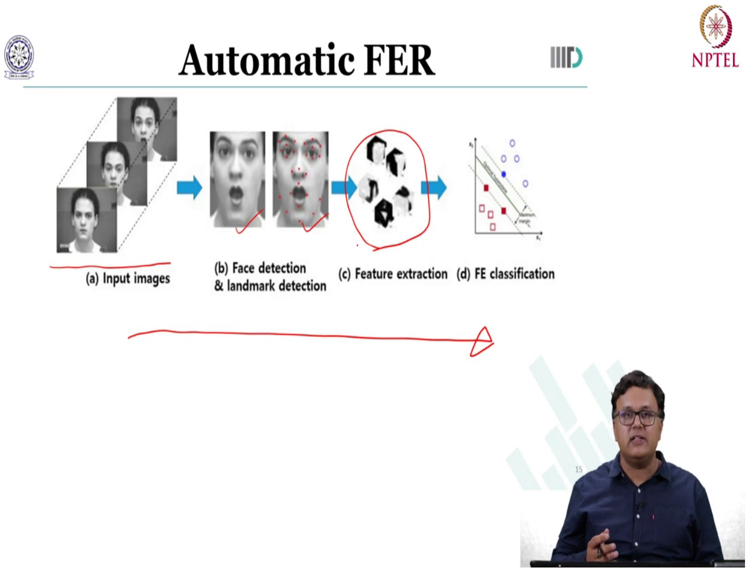
Week - 04
Lecture - 02
Automatic Facial Expression Recognition

Welcome to the 2nd lecture in Automatic Facial Expression Recognition as part of the Affective Computing Lecture Series. So, friends, in the last lecture we discussed about the components of typical automatic facial expression recognition pipeline. We also talked about the different variations with respect to the type of data.

You can have static facial expressions that is only one frame or you can have your dynamic facial expression where you have a series of frames coming in. Further, we also talked about the types of expressions essentially which tell you the label. You can have your macro labels which are your universal expressions, your angry, discussed, happiness, sadness and surprise and so forth.

And you can have micro expressions, the twitch of the eye, the subtle movement of the lip corner. Now, these are involuntary movements, but these are extremely important for understanding the effective state, the emotional state of a person. Then we talked about the facial action coding system where you have the action units, different parts of the muscles coming together to create a smile, to create an expression of surprise for this side ok.

(Refer Slide Time: 01:44)

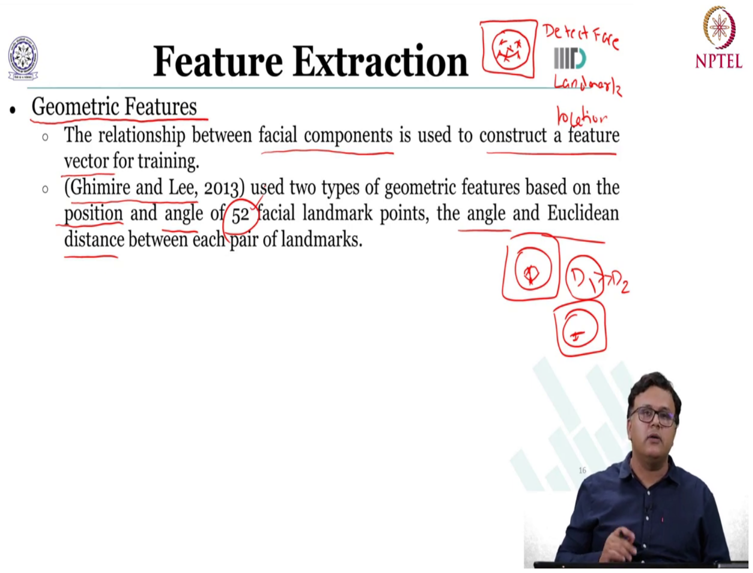


Automatic FER

Now, what you see here on the slide is your typical pipeline of a facial expression recognition system. You have the object detection where frames come in, you get the face you get the facial landmarks. And then as we discussed the last time, we are interested in extracting useful meaningful features, statistics around the facial region, the facial parts.

So, we are going to now deep dive into the facial expression features and see how these features are extracted, how they have been proposed in the past 20 odd years and what can be their advantage and disadvantages.

(Refer Slide Time: 02:21)



Feature Extraction

- Geometric Features
  - The relationship between facial components is used to construct a feature vector for training.
  - (Ghimire and Lee, 2013) used two types of geometric features based on the position and angle of 52 facial landmark points, the angle and Euclidean distance between each pair of landmarks.

So, we are going to start with the most simplest of a feature ok. So, let us say here I have a face, now in this frame I draw a smiley alright. So, assume that this is your face. Now, what I can do is I again detect the face using a object detector and then I find out the location of the parts which are the landmark locations.

Now, once I have these landmark locations. So, I am just you know pinpointing them here around the facial parts, I can extract what is referred to as geometric features. Now, as the name suggests, we are interested in extracting the facial geometry and this facial geometry is a factor of the facial points which you would be extracting using an object detector.

Now, clearly at least for macro expressions there is a strong relationship between the facial components and the feature way of construction ok. So, when you use these geometric features, you can understand if a person is smiling or is a person sad because the facial points are moving as such.

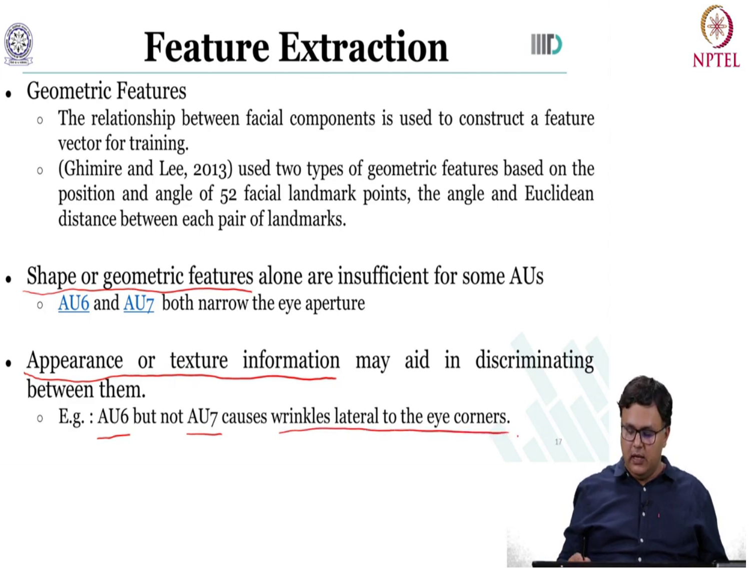
Now, in one of the very interesting works, Ghimire and Lee in 2013 they proposed that you can have two types of geometric features. One is based on the position whereas; the tip of my nose and the other is the angle. So, when I smile and you notice my face so, the angle between the outer corner of the lip and the lower center of the lip.

Now, you use the angle and the position and you can have different number of facial landmark points and you can create analysis of the expression based on the angle here and the Euclidean distance right. Now, this will be a feature. If for example, in a face the distance between the points here on the upper part of the lip and the lower part of the lip is high.

So, let us say this is D1 then if it is greater than another value D2 which of course, for the same face you are getting from a neutral expression. So, this distance then you can say well this is where the mouth is open. So, the chances are the person could be showing for example, a surprise or a happy expression right.

Now, notice in this work the authors had 52 facial landmark points you can have a larger number of points as well and these points are based on the type of object detector which you are going to use.

(Refer Slide Time: 05:06)



Feature Extraction

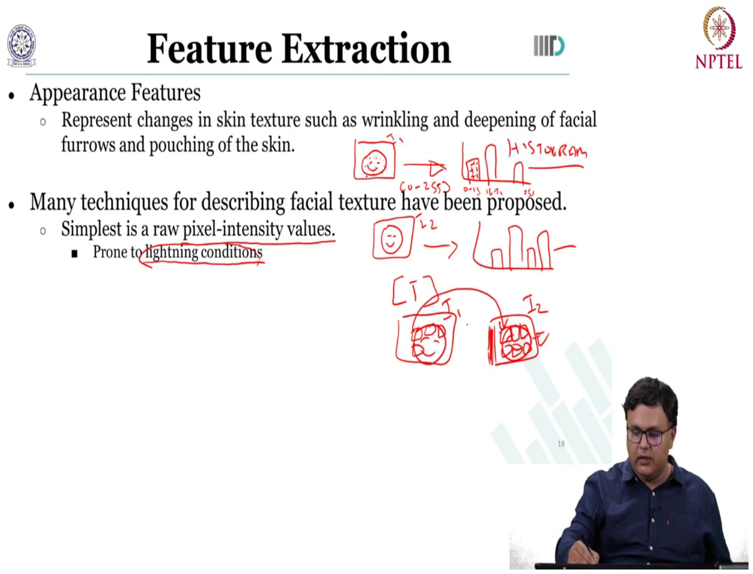
- Geometric Features
  - The relationship between facial components is used to construct a feature vector for training.
  - (Ghimire and Lee, 2013) used two types of geometric features based on the position and angle of 52 facial landmark points, the angle and Euclidean distance between each pair of landmarks.
- Shape or geometric features alone are insufficient for some AUs
  - AU6 and AU7 both narrow the eye aperture
- Appearance or texture information may aid in discriminating between them.
  - E.g. : AU6 but not AU7 causes wrinkles lateral to the eye corners.

It is noticed that shape or geometric features by themselves are not always sufficient. The reason is as follows you are trying to analyze the expression of a person and the person let us say gives a smirk notice my face right and I give a smirk. Now, this smirk if you have lesser number of points will not be appropriately captured by the landmarks points.

Another example is when you have action unit 6 and action unit 7 coming together then you see the narrowing of the eye aperture right. If you have lesser number of facial points then the geometric feature will have insufficient information. Therefore, in a large number of applications we use a combination of not just geometric features, but also feature extracted from appearance or texture.

Now, when you will add appearance or texture that is going to help you in discriminating these scenarios for example, where you have AU6, action unit 6 which is present, but action unit 7 is not present right and that is going to give you wrinkles for the eye corners right that is very vital information. Now, friends let us talk a bit about the appearance textures.

(Refer Slide Time: 06:38)



Feature Extraction

- Appearance Features
  - Represent changes in skin texture such as wrinkling and deepening of facial furrows and pouching of the skin.
- Many techniques for describing facial texture have been proposed.
  - Simplest is a raw pixel-intensity values.
    - Prone to lightning conditions

Now, appearance means essentially the skin component of your face right. So, we would like to have a texture feature descriptor which can analyze the different parts of a face such that you can have the information about the change in the skin when a facial muscle moves. Now, there are a large number of approaches which have been proposed in the literature for expression based features when you are analyzing the appearance ok.

The simplest base is well given a face you can extract the pixel intensity values ok. Now, what that will mean is let us say here you have a frame you create a face here ok you are going to analyze each and every pixel in your face and then let us say create a representation based on the pixel information.

So, you can create a histogram for example, you could say I am going to create a histogram which is going to have the range let us say its 0 to 255 is the intensity range. So, you are going to create n number of bins. So, 0 to 15 for example, then you go from 16 to 31 tell you know similar to 255 and you scan all the pixels and you keep on adding here ok keep on adding here you increase the frequency right.

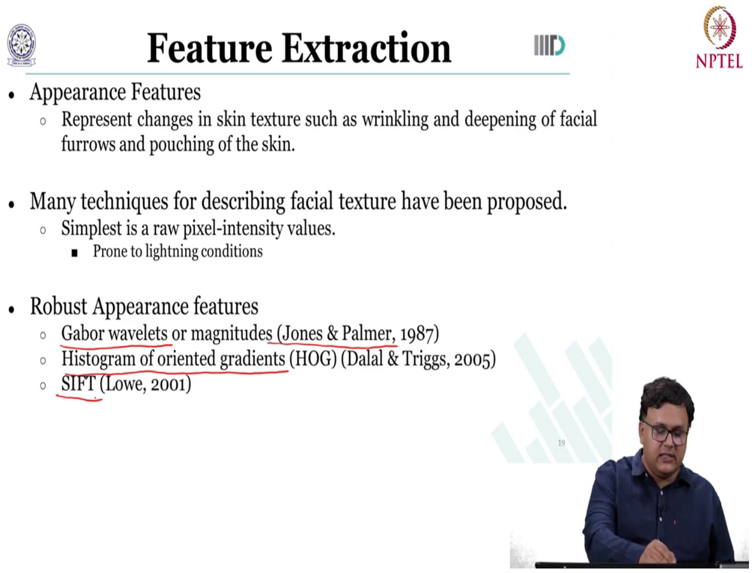
Now, that is another feature descriptor based on pixels. Now, a big limitation of pixel based appearance analysis is the weakness of pixel intensities to the change in illumination. So, when the lightning condition changes the pixel intensity changes and; that means, let us say when in image 1 same expression was there and then you had image 2 captured after a while with the subject showing the same expression.

But in a different lightning then the histogram which you would create based on the intensity that will be different. Now, histogram here histogram here notice they are from the same facial expression of the same person but are different because the lightning condition has changed.

Another issue is when you have effect such as translation now what; that means, in frame 1 you had your face this is your frame 1 and in frame 2 the same person moved a bit let us say by a few pixels. Now, the expression remained the same now since you are having this intensity based features some new information will be added here some information might go out we will not have always the same feature representation.

If you were to then divide the face into patches right local regions you wanted to compare region by region. So, that you have in detail and comparison of the facial regions that is a feature from here to feature from here the translation will have an effect here.

(Refer Slide Time: 10:05)



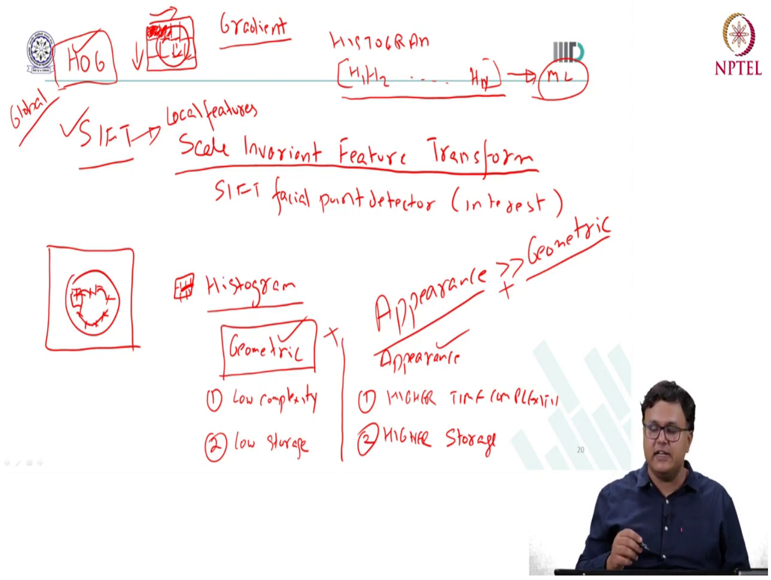
Feature Extraction

- Appearance Features
  - Represent changes in skin texture such as wrinkling and deepening of facial furrows and pouching of the skin.
- Many techniques for describing facial texture have been proposed.
  - Simplest is a raw pixel-intensity values.
    - Prone to lightning conditions
- Robust Appearance features
  - Gabor wavelets or magnitudes (Jones & Palmer, 1987)
  - Histogram of oriented gradients (HOG) (Dalal & Triggs, 2005)
  - SIFT (Lowe, 2001)

Therefore, in literature it was proposed that let us use orientation based gradient based information. Now, it has been shown that our eyes are also sensitive to change in the shape. Now, similarly for analyzing the change in the shape which is reminiscent to the facial expression you can use the gradient information ok.

Researchers have used the standard Gabor filters the Gabor wavelet filters to look at the different orientations the different changes in the facial expression when you apply different filters and then they will do let us say a histogram based representation or similar representation or the more popular ones for example, the histogram of oriented gradients and the scary variant feature transform. Now, let us discuss these two ok.

(Refer Slide Time: 11:06)



Now, your histogram of gradient for facial expression analysis will be as follows. So, you have a face which is detected using an object detector. What you do is you divide the face into non overlapping blocks for example, here I am dividing. Now, further what I am going to do is I am going to compute the edge map.

So, that what I get is the gradient. When I compute the change in x direction and y direction, I get information for the gradient. Now, what I will do is, I will create a histogram of gradient for each block here this is block 1. So, let us say you get H 1. Now, you append that with the histogram which you get from the second block and you then do it for let us say if they were N blocks.

Now, there are some post-processing operations which are done to increase the robustness of this feature to illumination ok. For example, normalizing it based on the variance and mean.

Now, this feature is an extremely powerful feature which when you will give to your machine learning algorithm should be able to differentiate the macro expression and micro expressions as well.

The other one friends which is extremely popular is your scale invariant feature transform. Now, this was proposed by Professor Lowe in the early 2000s. Now, scale invariant feature transform based facial expression recognition would mean here you have a phase. Now, you run the sift facial point detector which is essentially the interest point detector proposed for sift what that will give you is these points.

Now, these points are the important points where there is change happening in the texture in all the directions. Now, once you get these sift points very similar to how you did for your histogram of gradients you will take the region around a point let us say I got this interest point. So, this is the region around this point this is the I region.

You will create a histogram and this histogram will encode local information. So, small histograms coming together based on the orientations again right. So, you can either use your hog or sift or variants of these and then you can learn again a machine learning system. Now, it has been shown in a large number of studies that these appearance features they perform much better in terms of classification accuracy as compared to your geometric features.

Of course, one can do fusion as well that is we have a parent's feature and a geometric feature and you then combine them before you learn a machine learning classifier. Now, for geometric features the advantage is that they are extremely fast in computation. So, low complexity.

So, they have low time complexity and also, they have low storage requirement because all what you are doing is you have some facial points and on the basis of that you are doing some comparisons for example, computing the distance between facial points or computing the angles.

Now, compared to this the descriptors from appearance they have higher computational complexity because you are analyzing pixel by pixel. So, they are generally slower to compute higher time complexity and they have higher storage requirement. Therefore, depending on the type of application which you want to develop where facial expression is going to be extracted you can choose geometric appearance or combination of both.

Now, a very simple example which I would like to share friends is let us say you want to create a simple app in your phone which analyzes the images which are taken from a camera and then based on the facial expressions it clubs the images together ok. So, you join images together based on facial expressions and you can say well I have a feature in my app which can show the most happiest moments of a particular event ok.

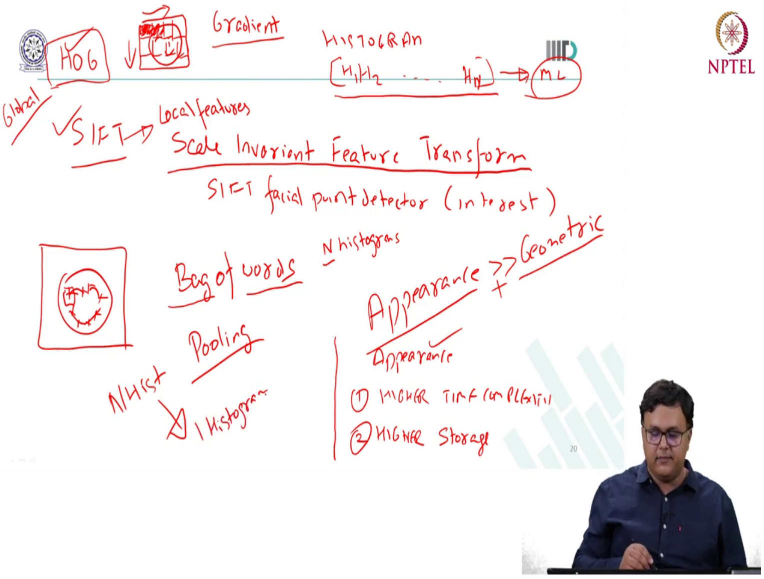
So, in that case if the faces are clear and closer to the camera; that means, they have a large face size in the images what can one do is you can have an object detector. And then you can extract the facial points and for things such as happy a geometric features could be good enough to increase the performance a bit more. For let us say those expressions which are not captured by the geometric feature one can have a histogram of gradient where you have less number of blocks.

So, as to have less time complexity. What this also means friends is when you were extracting appearance feature based on hog you were actually doing a global feature extraction for the face. When you were using your sift you were first getting the points the important interest points on the face and around each interest point you were extracting a histogram, which means we will have what is referred to as local features extracted.

Now, an obvious question arises if given a face I have n interest points around each interest point I compute an appearance feature descriptor based on sift for each image now I will have n sift feature descriptors. How do I learn a classifier? Because typically what you will have is each data point is represented by a histogram but. When you are doing sift you have a large number of histograms.

Therefore, when you have this kind of situation typically we have these pooling methods which are used for expression recognition and I will give you a quick example of one.

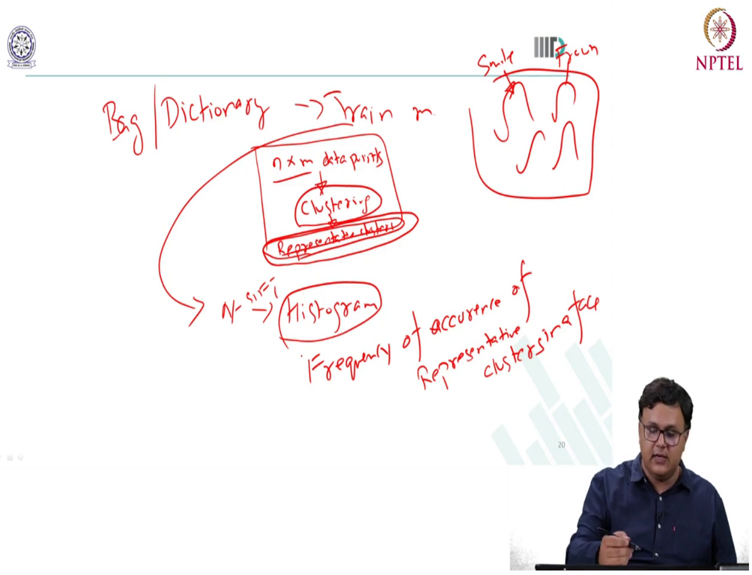
(Refer Slide Time: 19:19)



So, a simplest one is called your bag of words. Now, this concept is coming from the language domain what you are saying is well I have N histograms which are around N locations in my face. Let each histogram be one word and each face be a bag what I will do is I will create a pooling mechanism.

So, that these n histograms they can be created and compressed into one histogram, which I can further use to do training. Now, quickly how do we do that let us first erase ink on the slide.

(Refer Slide Time: 20:17)


Bag | Dictionary -> Train m
n x m data points
Clustering
Representative clusters
N-SIFT -> Histogram
Frequency of accurence of Representative clusters in a face
Smile
Frown
NPTEL
20

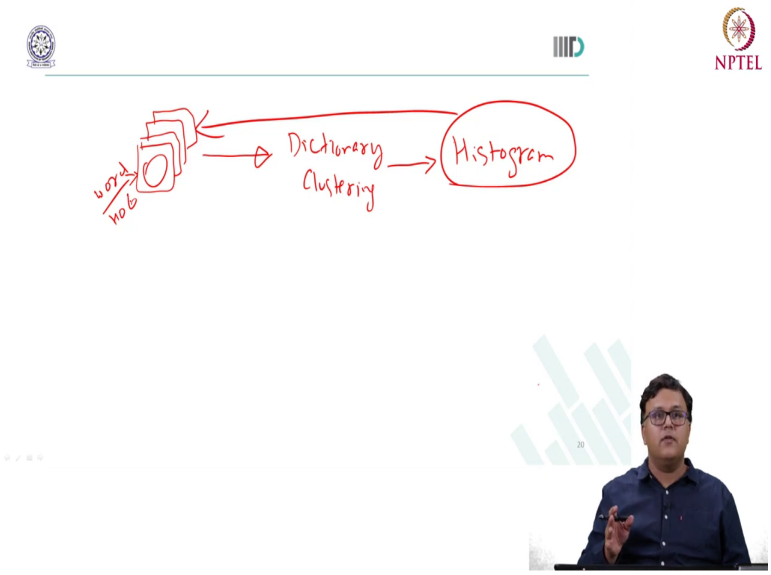
What you are saying is I will create a bag or a dictionary. So, I will take all the train samples and let us say I have m train samples basically m faces and I may have N interest points. So, I will have n into m data points I will do clustering and get what is referred to as representative clusters.

Now, what that should mean let us say in your faces you have a set of smiling faces, sad faces, neutral faces you extract these interest points you do clustering. So, you have your points you created these clusters maybe cluster 1 corresponds to a smile, cluster 2 corresponds to a frown. So, now, you have these representative clusters what you will do is you will take one sample at a time. So, you take one data point which contains N interest points.

And you will create a histogram by looking at the frequency of occurrence of representative clusters in a face you compute the distance between the sift histogram for each point with all the features of clusters and you assign to the nearest neighbour and that will give you the histogram.

Now, this solves a lot of problems not only friends now I can have a method for N different facial points for a face, but if instead of a face a single frame I was given a video if I was given a video then as well, I can do pulling using back of words right.

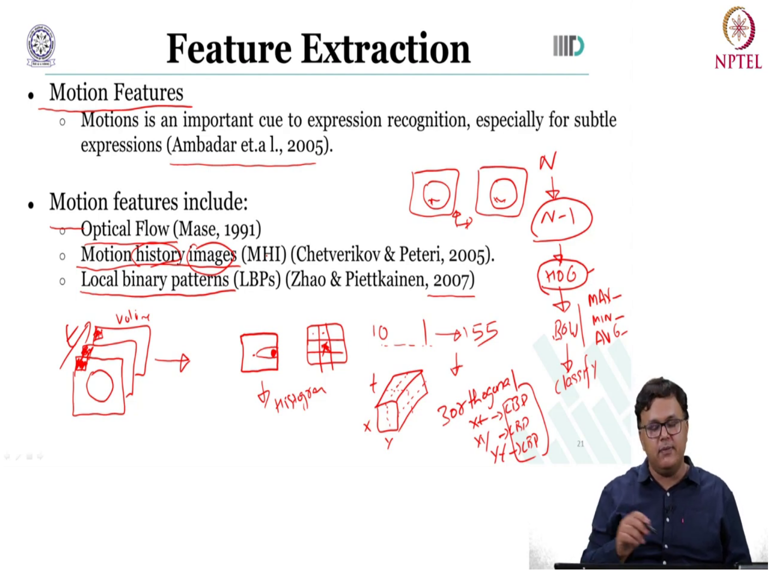
(Refer Slide Time: 22:40)



So, you have frame one you have frame two you ran your object detector you got the face you got the landmark points and let us say you extracted global or local features right. So, what you will do is you will create a dictionary using let us say clustering. So, you could say well one representation is one frame is a word and this is represented by let us say a hog based feature.

You do dictionary based clustering then you do the histogram process creation which essentially referred to as your vector quantization now you will have a single histogram for E1 video ok. So, this could be a face across all the frames and this way you can have facial expression recognition for dynamic where you have a series of frames coming in.

(Refer Slide Time: 23:35)



Feature Extraction

- Motion Features
  - Motions is an important cue to expression recognition, especially for subtle expressions (Ambadar et.a l., 2005).
- Motion features include:
  - Optical Flow (Mase, 1991)
  - Motion history images (MHI) (Chetverikov & Peteri, 2005).
  - Local binary patterns (LBPs) (Zhao & Piettkainen, 2007)

Now, we also have another way of extracting motion features see when I say you can create a back of words based representation for a video that is going to analyze each frame as a word I am not learning the relationship between each sequential frame you started smiling right the onset apex offset. So, every frame one has a relation with frame two.

Therefore, an ideal feature representation should be the one which not only looks at the spatial information for example, when you apply hog to a frame you extract spatial information, but also looks at the time series right. The changes which are happening across time because facial expressions are dynamic in nature.

Therefore, motion based features are extensively used in a work by Ambadar and others they showed that you know when you add motion based information into your features you can do better facial expression recognition. There are a large number of methods let us look at the most prominent and commonly used ones. So, the first one is optical flow.

So, optical flow is a traditional computer vision based technique where you say well given a series of frames I want to know where my pixel from frame one moved to in frame two. In context of your faces that will mean you had smile starting in second frame you had the same face and let us smile increasing a bit.

Now this frame at the corner this pixel at the corner that has moved a bit in the next frame right. So, what is that flow which is the velocity and the direction in which a pixel has moved. So, once I compute this optical flow then I can have different type of pooling methods again similar to bag of words.

So, friends what happens is you have N frames let us say as input you will get N minus 1 optical flow frames which tells you know how pixel is moving from first frame to other frame. Now you can extract any feature you can say well I want to do a histogram of gradient again to each optical flow based frame and then a bag of words and then classify.

So, now what is happening we took the face as an input we extracted the flow it told me how are the points moving across time and then we looked at the change in spatial domain as well. So, each optical flow frame now has a histogram representing the orientations of the gradients you can do some pooling because you have a video you know bag of words is one could say I want to go even more simpler because I do not want to spend more computation cycles right.

So, I could say well then just compute the max, min and let us say average across all the hog frames you get three frames you can flatten them and then that is your feature vector. Friends other one is motion history image this is a bit older one as well and it has been used more for human action recognition and has been shown to be useful for facial expressions as well.

What you are saying is I want to create one single image from a series of frames and these frames are going to be flattened into one frame wherein, I am going to look at the change which has happened in a particular location. So, essentially you know this one single image is going to create a contain the history of motion of the pixel across that location the other one which is extremely popular and very effective for dynamic facial expression recognition is your local binary pattern.

Now Zhao and Piettkainen in proposed this method in 2007 and let us see what this method is. So, friends what you are saying is you have a series of frames again of a single person. So, we are assuming always we have a single person. Now what you do is you would take one block of the frame here and similarly same size block in the same location, but now in the second frame.

So, from this volume you would extract a sub volume. So, all these non-overlapping first blocks. What you will do next is you compute the local binary pattern what that local binary pattern says in its simplest form. Let us say I am right now analyzing the location of the tip of the eye. So, this is the outer tip of the eye.

You will take one pixel so, this pixel and what you do is you compare this pixel with its local neighbour. So, this is the pixel this is the tip of the eye you compare this with the neighbouring intensities now let us say the intensity here if that is larger than my neighbour.

Then I would say this is a one now I compare the intensity of this pixel with this one if this was let us say less I will say it is a 0. So, now, if it is larger it is a one if it is smaller than the neighbour it is a 0 this way I will get a 8-bit code ok. Now you compute the required the decimal point for this now let us say 55 just as an example friends.

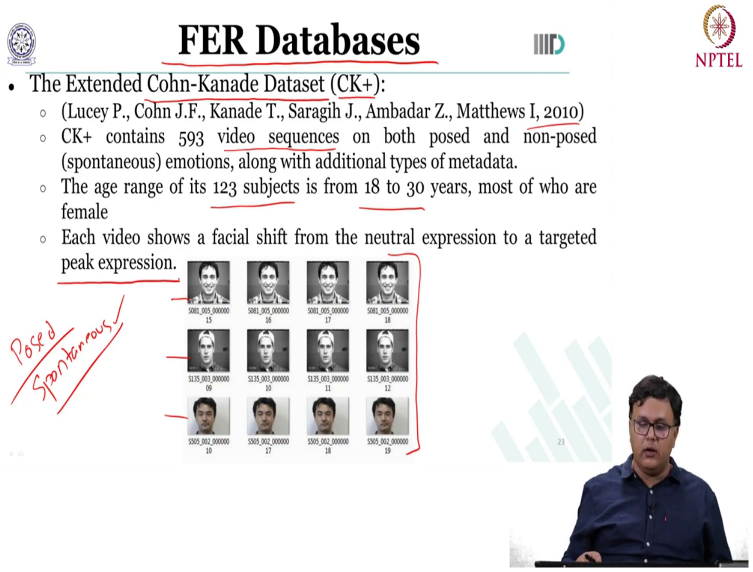
And then you will create a histogram of all the points right you have created these binary code for all the points now this should give you one histogram ok. Now this is taking care of the spatial part now Zhao and Piettkainen said well. That is for the spatial part now let us do something for the temporal part as well.

So, what I am going to do is let us say I take again this non overlapping block first block of my first frame then I take for a second and third frame and I have the sub volume ok. So, I am going to draw the sub volume. So, this is x this is y this is t this is at well. Let us divide the volume into 3 orthogonal planes what; that means, is let us say I have this frame here which is your x t then I have your frame in the here in the center which is your xy and here I have a frame which is your yt.

So, we have three frames from each you compute one LBP. And then you combine this and this is now your spatiotemporal analysis of the video using a local binary pattern representation local binary pattern is also used for texture analysis in computer vision. What that means, in context of phases is for applications such as micro expressions these have been observed to be fairly discriminative.

Because you are analyzing the pattern in the texture pixel by pixel and are encoding the difference of that pixel with its local neighbourhood in both space and time. So, once you have any of these features friends you will then learn again the classifier the same drill we have been discussing about.

(Refer Slide Time: 32:49)



FER Databases

- The Extended Cohn-Kanade Dataset (CK+):
  - (Lucey P., Cohn J.F., Kanade T., Saragih J., Ambadar Z., Matthews I, 2010)
  - CK+ contains 593 video sequences on both posed and non-posed (spontaneous) emotions, along with additional types of metadata.
  - The age range of its 123 subjects is from 18 to 30 years, most of who are female
  - Each video shows a facial shift from the neutral expression to a targeted peak expression.

Posed

Spontaneous

Now these are some of the methods for your features which we are going to extract another important aspect here is the data right. Affective computing is a data driven area. So, let us look at some of the facial expression recognition databases which are extremely popular in the community and have very different purposes which helps us in solving different problems.

So, we have the Cohn-Kanade dataset from the robotic institutes at CMU where you have CK dataset and then later you know the CK plus dataset which was collected in 2010 and what this dataset has is it has video sequences where a person was asked to come and sit in front of the camera, now if you observe me guys I am looking right into the camera and the person was asked to smile then the person was asked to show sad expression.

Now, if you notice what has happened is my head did not move, but the expression change right now researchers wanted to have pure expression movement in these samples and that is how you get these points data points where you have the video where the expression is changing further, there are over 100 subjects there is a diverse age range notice this is actually not containing much older subjects.

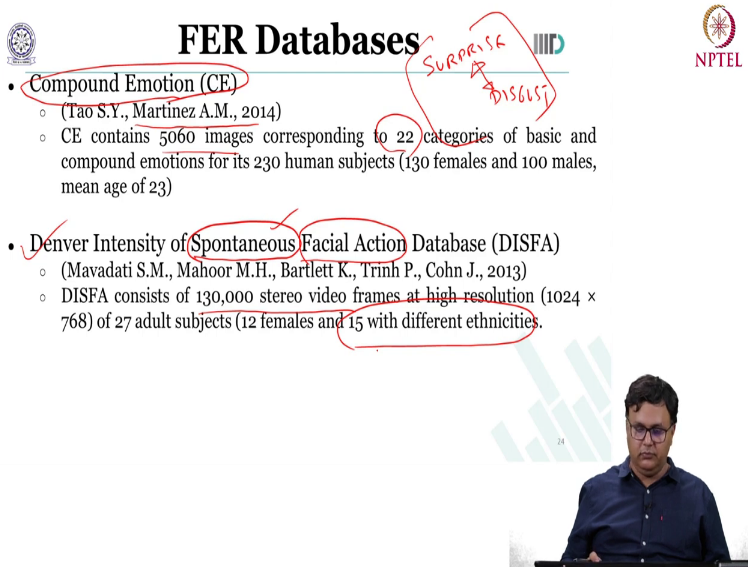
This is mainly you know graduate students and faculty around the campus now CK and CK plus dataset have been the main bread winners with respect to the datasets which are used for learning clearly, they have added to much progress in the area, but as I explained to you the method with which the data has been recorded; that means, that the expressions are not spontaneous right.

So, that brings us to the discussion of what is spontaneous and what is posed expression right. Posed is you looked into the camera and then you smiled spontaneous is let us say you were having conversation you were watching a movie there was a joke you really appreciated you liked the joke.

And then you smiled right. So, that is spontaneous in the real world we have more spontaneous expressions, but certainly you have posed as well. It is similarly you can say you have a group of friends who are now posing in front of the camera and they are saying cheese right when you say cheese it is a posed expression right.

Now, mainly these are posed expressions in the case of the Cohn-Kanade dataset another aspect to notice we discussed that in the case of static facial expression recognition we are mainly looking at the peak expression. So, researchers here also labelled the peak expression which in other words means the expressions highest intensity which is the apex ok.

(Refer Slide Time: 36:16)



FER Databases

- Compound Emotion (CE)
  - (Tao S.Y., Martinez A.M., 2014)
  - CE contains 5060 images corresponding to 22 categories of basic and compound emotions for its 230 human subjects (130 females and 100 males, mean age of 23)
- Denver Intensity of Spontaneous Facial Action Database (DISFA)
  - (Mavadati S.M., Mahoor M.H., Bartlett K., Trinh P., Cohn J., 2013)
  - DISFA consists of 130,000 stereo video frames at high resolution (1024 × 768) of 27 adult subjects (12 females and 15 with different ethnicities.

SURPRISE & DISGUST

Now researchers moved on from this and then they started looking at different aspects with respect to the recording environment with respect to the age range the cultural variability and also the labels right. In the case of your CK dataset you had the compound emotions and the sorry the universal emotions and the facial action coding system facts action unit labelling. Another work from Ohio State University in 2014 proposed a dataset for compound emotions.

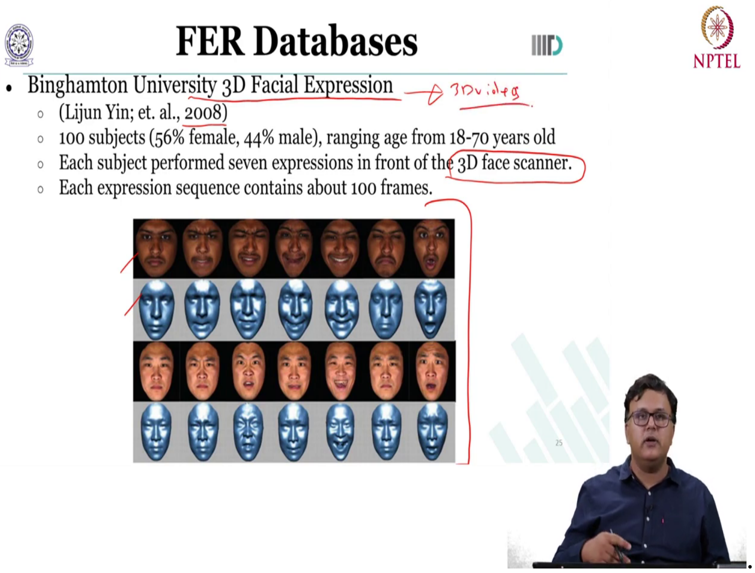
They said well universal emotions are too less right and they are not universal as well we have finer states in between. So, for example, you have surprise and you have discussed right these are two different expressions in the case of universal emotions you could have a scenario where you could have a person who is disgusted and surprised right so, the emotion in terms of the expression that is compounding.

So, Martinez and others they proposed this dataset containing 5060 what images from 230 subjects and had 22 categories of basic and compound emotions, which means now from the perspective of facial expression recognition system you have a 22 class problem and; that means, you can have more fine grained information about the state of the user.

Then friends we also have the DISFA Dataset from the University of Denver. So, it stands for the Denver Intensity of Spontaneous Facial Action datasets. Now, as the name suggests this is spontaneous; that means, the users were shown let us say a series of videos could have positive emotion or negative emotions and the expectation is that some emotion will be induced in the viewer.

And that will be shown in terms of facial expressions right. So, spontaneous expressions and then there was human labelling for facial actions. So, you have now 3D data at a high resolution and different ethnicities as well right. So, there's another type of variability which is coming in this dataset.

(Refer Slide Time: 38:42)



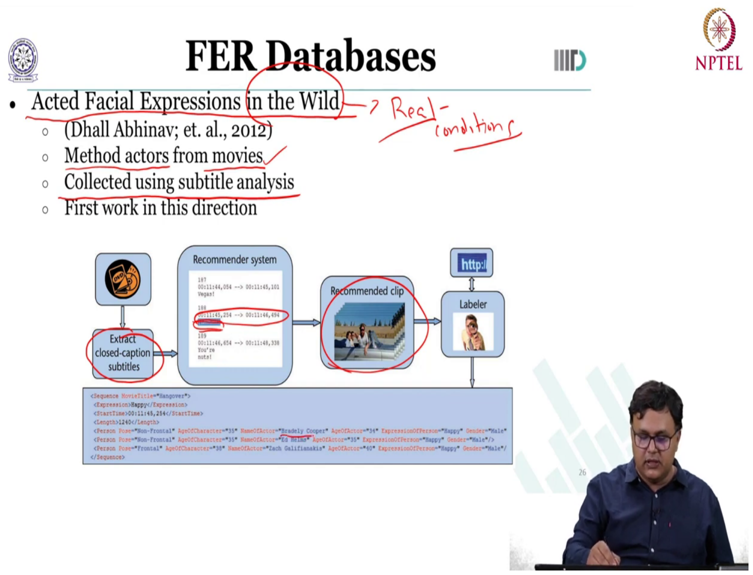
FER Databases

- Binghamton University 3D Facial Expression → 3D videos
  - (Lijun Yin; et. al., 2008)
  - 100 subjects (56% female, 44% male), ranging age from 18-70 years old
  - Each subject performed seven expressions in front of the 3D face scanner.
  - Each expression sequence contains about 100 frames.

Now, all the examples which I have been talking till now have been about 2D data, but we also have 3 dimensional data friends right. So, you can have a 3 dimension data either it can be captured from 2 cameras or it could be 3D data which is synthetic right. So, Lijun and others they propose a dataset called the Binghamton University 3D facial expression dataset which samples are shown here.

So, what you have here is the texture and the shape right and this again was created in a bit different manner ok. So, they use a 3D face scanner and then they captured the faces of the participants. In the next version they actually captured 3D videos right now when you capture 3D videos again one is posed and then you can go to the next step which they did. Wherein they showed the videos stimuli to participants and then they got 3D videos and then they got spontaneous 3D facial expressions.

(Refer Slide Time: 39:51)



FER Databases

- Acted Facial Expressions in the Wild → Real-conditions
  - (Dhall Abhinav; et. al., 2012)
  - Method actors from movies ✓
  - Collected using subtitle analysis
  - First work in this direction

Extract closed-caption subtitles

Recommender system

```
187
00:11:44,054 --> 00:11:45,101
Vegas!

188
00:11:45,254 --> 00:11:46,494

189
00:11:46,654 --> 00:11:48,338
You're
nuts!
```

Recommended clip

Labeler

```
<Sequence MovieTitle="Hangover">
 <Expression>Happy</Expression>
 <StartTime>00:11:45,254</StartTime>
 <Length>1240</Length>
 <Person Pose="Non-Frontal" AgeOfCharacter="35" NameOfActor="Bradely Cooper" AgeOfActor="34" ExpressionOfPerson="Happy" Gender="Male"
 <Person Pose="Non-Frontal" AgeOfCharacter="35" NameOfActor="Ed Helms" AgeOfActor="35" ExpressionOfPerson="Happy" Gender="Male"/>
 <Person Pose="Frontal" AgeOfCharacter="38" NameOfActor="Zach Galifianakis" AgeOfActor="40" ExpressionOfPerson="Happy" Gender="Male"/
</Sequence>
```

Now this is one of the works which I was involved in very early on in my PhD 10 years ago. So, the datasets which you have seen till now they have been focusing on facial expressions, posed or spontaneous, but the assumption is that the subject the user is always inside the laboratory with good illumination and mostly facing the camera so, looking into the camera as right now I am doing.

But this is not how exactly always the data will be when you capture the world around you many a times you will see that the pose of the person that is the head with respect to the camera would be non-frontal. For example, I am looking towards my left right now in this case the camera will only be captured in the right hand side of my face.

So, we have to introduce these kind of data such that the methods can be tested for these real world conditions. Now, along with the pose there are other aspects to real world conditions as well for example, there is occlusion there is presence of multiple subjects there is presence of missing data as well. Now to progress work in this area, we proposed a dataset called acted facial expressions in the wild.

Now in the wild here friend means this is a representative of real world conditions. Now, what we did was we chose popular movies which had method actors now method actors are these high quality actors who would generally show expressions which are closer to spontaneous expressions, but since we are getting data from movies the environment in which the people are that is very varied right now what we did was we said well we would get these movie data.

But we need a method to extract those short video segments where a person or a group of persons are showing the expression right could be happy sad and angry and so forth. So, what did we do we extracted closed caption subtitles now friends closed caption subtitles are those subtitles which you would have generally observed when you watch these foreign materials which is dubbed into let us say another language and has square brackets around it.

Now, these words for example, you know here you see laugh these generally will convey the emotion or something which is in the scene and these are for people with hearing impairment. So, we did this simple analysis of the text and we chose those segments which had keywords related to affect an emotion and then gave us this kind of data which was later either label connected.

Because you know you can always say that the word is related more to happy or not to sad and then use that label and along with this we also had the context. Now the context is as follows who is the person what is the age and what is the gender. Now we can use this information for facial expression recognition training and have more accurate prediction for let us say a better model, which predicts facial expression for older population a more accurate model for facial expression prediction in children right.

So, this brings us to the end of this lecture we have discussed the process of feature extraction we started with the features for frames geometric features. Then we moved on to appearance based features and then spatiotemporal features and later we discussed the different facial expression datasets and their attributes. For example, recorded in lab or recorded outside posed expression or spontaneous expression and then the different aspects such as was a 2D camera used or a 3D scanner was used.

Thank you.