

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

Week - 04
Lecture - 03
Automatic Facial Expression Recognition Group Emotions

Welcome back, I am Abhinav Dhall and we are going to discuss the third part of Automatic Facial Expression Recognition, which is in the series of the Affective Computing course. So, friends we have been talking about the different techniques in facial expression recognition from the perspective of static versus dynamic facial expressions, posed versus spontaneous facial expressions.

The different features which we extract and the different datasets which have been proposed in the effective computing community. Now, we are going to discuss about a newer dimension in automatic facial expression recognition. So, if you notice on the slide here.

(Refer Slide Time: 01:05)

The slide features a title 'Facial Expression Recognition' with a logo on the left and the NPTEL logo on the right. It contains two images: a single subject labeled 'CK+' on the left and a group of people labeled 'Google Images' on the right. A blue arrow points from the single subject to the group. Handwritten red annotations include: 'posed' with a checkmark next to the CK+ image; 'cohesion of a group' written below the CK+ image; 'overall perceived emotions of a group' written below the Google Images image; 'universal' and 'Valence Arousal' written above the Google Images image; and a large green watermark 'NPTEL' behind the speaker's video feed in the bottom right corner.

So, what we have here is a frame from the extended Cohn-Kanade Plus dataset. As you notice here, there is a one subject, this subject is looking directly into the camera and there is a posed expression, ok. So, this is a posed expression. Now, we discussed earlier that we would have spontaneous expressions in the world as well, right that is how we communicate through nonverbal communication.

What we have also seen is given that there is so much data which is now available on social media and also that there are these circumstances where we are not alone, we have a group of people around us. So, in that case, there can be scenarios like this, where you have a group of people and a camera is recording either an image or a series of images or video.

And what we want to understand is that, what is the overall perceived emotion of this group, right. Why do we want to do this? Well, of course, I told you one reason already, there is

abundant amount of data available and then we also see that in task such as understanding the cohesion of a group that is an indicator of let us say the unity within a group, the task where we want to understand how a group of people are performing.

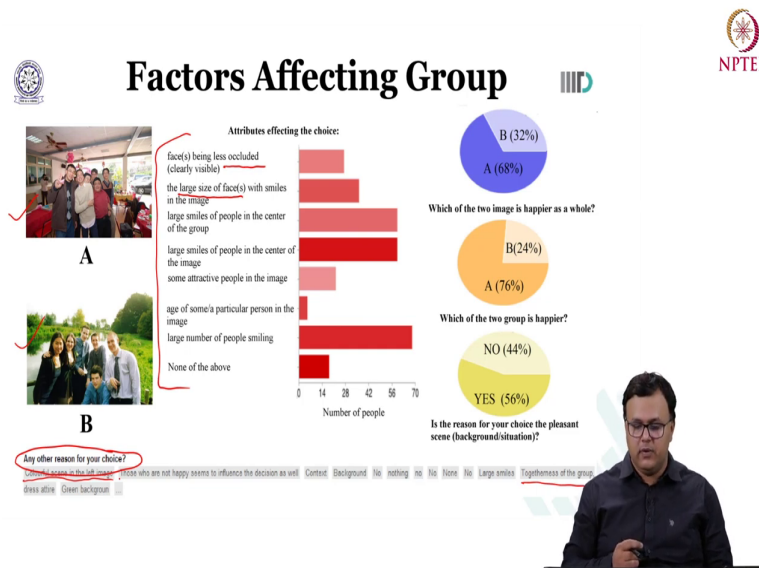
So, in that particular case, the facial expressions along with other cues gives us a vital cue. Now, as compared to traditional facial expression recognition where we have one subject, in this case there are a large number of factors which will affect the perception of a person when they look at let us say photograph of a group of people.

Now, we would like to map these factors onto a computational model so that we can do overall facial expression recognition for a group. Now, the cohesion of course, is how can we represent the emotion of a group, right. So, in this similar to how we did for single subject based facial expression recognition, we can use the universal category.

So, that is your angry, sad disguised, happy neutral and so forth. You can also use here the valence and arousal continuous emotions, right. So, we have seen in the earlier lectures, we can use this valence and arousal to indicate the emotion and its strength positive or negative.

Now, if you have a video of a group of people, you can then indicate that what is the primary emotion and then what is the intensity, right. Now, we will go through this part trying to understand first what is going to be the contributing factor when someone looks at a photograph, why will they will say for example, a group of people look happy or they are in a professional setting. So, maybe they are showing a neutral expression.

(Refer Slide Time: 04:30)




So, we did a survey, ok. Now, this is one of our earlier works where we asked users a question where we said well you have 2 images, can you compare these 2 images and tell us that which group of people is happier. Now, once you have chosen the choice, there are a set of questions which are derived from works in psychology which have already studied how humans are perceiving the groups and what is their choice based on these factors, right.

For example, you could say you chose image A having better higher happiness as compared to B, is it that because the faces are less occluded or the faces in a particular group have larger size or maybe some people have particular attributes which are affecting your judgment. Based on that, we did some analysis and we found out for example, there are a set of contributing factors.


The example is larger faces, more smiling faces, they are getting more weightage. So, in a way when you look at a group of people and let us say this is a large group of people, you do not need to look at each and every face in order to tell me what is the overall group emotion, right. You can look at a few people and that can give you know a correct understanding of the perceived emotion.

Now, we also asked our participants that what are the reasons which other than the ones asked based on face size and others which made you choose a particular image. So, if you notice they say there is more togetherness in this group. So, I am rating the happiness of this group higher than the other one. There is a more colorful scene in the image. So, that will mean perhaps you know the event in which these people are that is more positive.



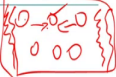
(Refer Slide Time: 06:44)



Factors Affecting Group



- Top-down affect: Neighbours, scene, group
- Bottom-up affect: Individual expression; Attributes – attractiveness, spontaneous, age, gender, large faces, centre faces, occlusion etc.
- Bar's scene context model:
 - low-resolution holistic representation, similar to scene descriptor
 - detailed object-level representation, face analysis



Now, moving forward in this, we group these factors which are affecting the perception of a group of few people into a top down effect and bottom up effect. Now, friends, what is top down effect? Top down effect is here you have a group of people and they are at a particular place. Let us say they are at a park.

Then what is the effect of the neighborhood of each person on another person? For example, for this subject, what is the effect on his or her expression based on with whom they are standing and where they are standing in the group? So, typically the structure of a group also tells you about the social setting of that group.

For example, if you are in a formal setting in office, then it is possible that the senior person could be in the center of the group. If you are with friends, maybe this condition will not hold. The other is the scene. What are the things which are in the background? Right, maybe if there are balloons, it can indicate it is actually some celebration going on or if it is actually, you know some chairs and people are wearing robes, then maybe a congregation is going on right. So, the clothes of the people and their body pose.


Now, the second category is the bottom up effect. This is essentially saying what is the individual expression and what is that contribution of this individuals expression towards the overall group promotion. And there are things such as attractiveness, the spontaneous expression, age, gender, large faces, center faces, clearer faces and so forth. So, here we are looking at each person at a time from a computers perspective.

Now, this is also somewhat in sync with bar scene context model which proposed that when we look at a particular scene, we would do a very low resolution quick scan of that scene. So, gives us the holistic picture of what you have in front of you. Once you have done that, we will pinpoint a region of interest and we will fixate on that. What; that means, is you are looking let us say at scenery.


Now, you will do a quick scan, look at the low resolution version of the scenery and then maybe fixate on the house which is there just beside the mountain because that is what is of


interest to you. So, in that way, first you are looking at the overall group, then you are looking at the individual people ok.

(Refer Slide Time: 09:38)



Group Representation





Top-down + Bottom-up

① Object detector (faces) →
 Group Image → Fully connected graph
 $G = (V, E)$
 $w(e) = \text{Distance to faces}$


② Min-Spanning tree → Prim

③ After H_i → $GEM = \frac{\sum H_i \cdot d_i}{n}$

$d_i = \|g - l_i\|$

$\theta_i = \frac{S_i}{\text{avg}(S_{\text{group}})}$

Latent Dirichlet



Now, there are several ways in which one could model these attributes into a facial expression recognition system. We are going to discuss one such model. So, let us say we have an input image. In this image, we have six subjects. Now, what we can say is we are interested in the top down representation, those effects and the bottom up right. For me to let us say first look at top down, I need to have a method of modeling the group.

So, typical pipeline which we have discussed till now for facial expression recognition, friends we use an object detector right. In this case when you will use the object detector for detecting the faces, it will give you six faces in this image. Now, once you have it, what you could say is well let my group of people in image be a fully connected graph ok.

So, you have a fully connected graph here. In this fully connected graph, your vertices V are the faces and the edges are representing the link between two people. For example, in this image F_1 and F_2 are linked, so, this is an edge. The weight of an edge is essentially let us say the distance between two faces.

Now, the distance between these two faces can be calculated based on the face location. So, you can take the center of the two faces and let us say calculate the Euclidean distance. So, now let us say the Euclidean distance between these two faces will be the weight of the edge between these two vertices two faces.

Now, top down was based on where are people standing, with whom they are standing and then what is the effect of their relative location onto the neighbor, onto themselves which is affecting their expression and that in of course, in the end is giving the impression of the overall groups perceived emotion. So, what we do is after we have a fully connected graph, we say well let us compute a min-spantree ok.

So, for example, you can use the Prim's mins span computation algorithm and that is going to tell you who is the neighbor of whom, right? So, you have a fully connected graph, you computed a min-spantree and now it actually gives you the shortest path and in the classic sense of this group as a fully connected graph it is telling me now what is the shortest path.

So, in a way it could give you for example, if F_1 was your first face, then F_1 is let us say link to F_2 , it is F_3 , then F_4 from F_4 you can reach maybe to F_5 and then you can reach to F_6 just as an example. Now, what; that means, is with this min-spantree F_2 's two neighbors are now clear to me. So, F_2 has two immediate neighbors, right?

So, I can now calculate more information about the relative location, right? Just as an example what I can do is I can say well the relative face size of a person with respect to his or her neighbor is a rough indication of where that person could be standing in a group, right? If that is the case, I can do something like this I can say; well, if θ_i is the relative face

size of the i -th subject in my group then that can be calculated as the face size which I can get from the face detection which I have done.

So, let us say that is S of i divided by the average of the size of the neighborhood, ok. So, I can get this from F_1 , F_2 and F_3 and this simply is a very rough indication of where other people are. Similarly, I can compute the distance of a particular person let us say F_1 from let us say the centroid of the group. Earlier I gave you an example, right? When you are in a formal setting generally seniors are in the middle of the image.

Same goes of family photographs as well, you may have noticed you know the grandfather, grandmother that could they would be sitting in the center, the elder ones would be on the sides children could be in the lap, right and so forth. So, once distance from the centroid can indicate how much is their weightage in that social setting.

In vision, computer vision it is also referred to as you know finding a very important person. So, you have a group of people who is the important person in this image, once you know that that gives you a fair bit of idea about the social setting of that group where they are what they could be doing and what could be the perceived emotion, right?

So, if that is the case, I can do it this way I can say the distance is actually based on the centroid of the group and then you can subtract it by the location which you get for the face. So, let us say the location is l for a group i and this is group so, you can compute this, ok. So, larger the distance, you can then penalize the expression.

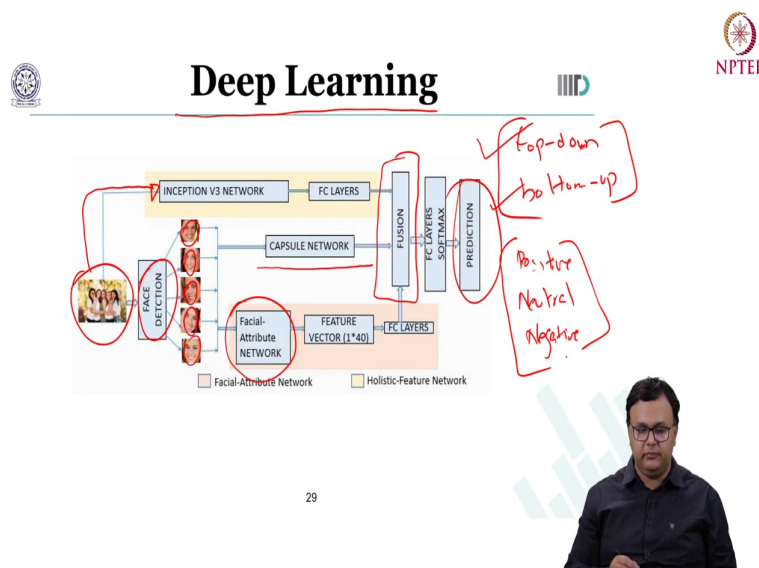
Now, similarly friends you can look at the bottom effect, right? I will give you an example. You can do AFER. So, that is your Automatic Facial Expression Recognition and you can compute of particular face let us say happiness, right? So, H of i will tell you how much happy of particular faces, right? Now, you have this information you can use d of i and θ_i as weights and apply it to the expression of one person.

So, you can down weight or increase their value based on their social significance in that group. Based on that trivial model of group expression let us call it group expression model

gem could be simply you know submitting the expression intensity which is weighted by θ_i and d_i and then you have the total number of subjects n in that group.

Of course, this is the simplest representation which you could have. You can have more factors added into it, but for example, let us say using machine learning techniques such as Latent Dirichlet allocation, right? LDA's. So, you can model the relationship between the faces, the group members wherein the word would be the face, similar representation. One could also use a bag of words with framework as we discussed in the last lecture for the dynamic facial expressions.

(Refer Slide Time: 17:15)



Now, these were techniques whereby you are using these handcrafted features. You could also use a deep learning based technique for group emotion, right? Here is an example. So, here you have an input image and again what you want is top down features to be taken care



of and your bottom up features, right? So, these are the ones which we understood from surveys.


So, this pipeline takes the whole image as an input into a network. So, in this case this is an inception V3 network and this is actually taking care of the top down, right? So, you are looking at the whole scene. Further what you do is you detect the location of the faces and for each face now you are using a network. In this case it is a capsule network; you can use a configuration neural network as well.


Now, this is your bottom up analysis, you are looking at the facial information here, ok? To add more context, for example, where people are, right? You will see for example, people could be wearing birthday hats only in a birthday social scenario, but not in offices, right?


So, you can have this facial attributes computed for each face and all this can be fused together. And this will be the prediction, right? So, you can do a prediction here, maybe let us say 3 class prediction which simply tells you if the perceived emotion is positive, neutral or negative and this is just a rough step on your valence axis, ok.

(Refer Slide Time: 19:00)

 **Group-Level Emotion** 







Now, I will play a video and this is an example of looking at the faces and the whole of the scene as well. Now, in this case you will notice that the facial expressions, the body gestures tell that the overall group emotion is not so positive, right? And here you have, you know, let us go back to the video, let us play it again.

And here on the top, right, you have the overall emotion which is negative and the bars are indicating the strength. So, higher score would mean that people are looking more negative as you can clearly tell from the expression of the group of people here. Now, in this very example, if I was to move on a bit, you will notice that there are a positive sample coming in as well. Let us wait for it to come.

(Refer Slide Time: 20:07)






Group-Level Emotion





Now; obviously, one could say that, you know, for example, in this group.


(Refer Slide Time: 20:13)

 **Group-Level Emotion** 









One can look at not just the expression, but the body gesture is well-trained.


(Refer Slide Time: 20:16)



Group-Level Emotion



(Refer Slide Time: 20:17)



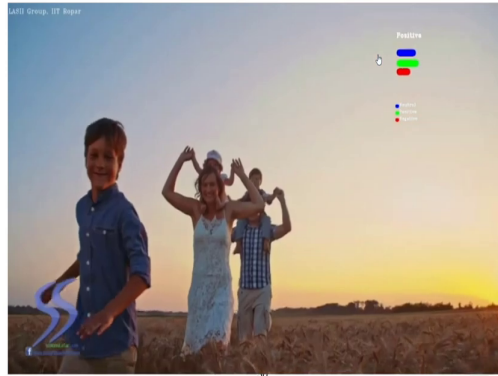
The slide features a central video frame showing a group of people in a field. A 'Positivity' bar chart is overlaid on the video, with a green bar indicating a high level of positivity. The chart has a legend with four colored squares: blue, green, red, and yellow. The video frame also contains a 'LIFE Group: 03: 04: 05' label in the top left corner and a '30' label in the bottom right corner. The slide is titled 'Group-Level Emotion' and includes logos for IITD and NPTEL.

The body gesture of the group of people can tell you, for example, they are, you know, having a positive emotion right now. So, these are very powerful cues which we can extract and later embed into a large number of systems.

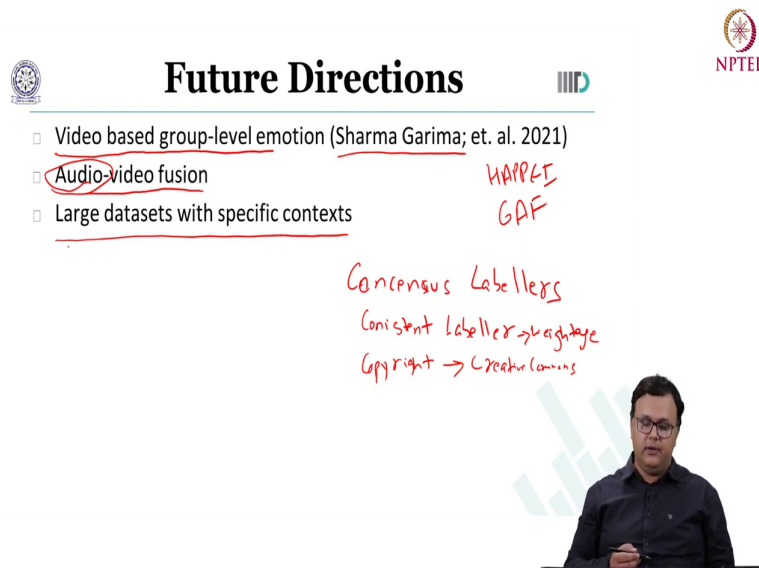
(Refer Slide Time: 20:26)



Group-Level Emotion



(Refer Slide Time: 20:29)



The slide is titled "Future Directions" and features a list of three items:

- Video based group-level emotion (Sharma Garima; et. al. 2021)
- Audio-video fusion
- Large datasets with specific contexts

Handwritten notes in red ink are present:

- Next to "Audio-video fusion": HAPPY, GAF
- Below the list: Consensus Labellers
- Below that: Consistent Labeller → Advantage
- Below that: Copyright → Creator/Owners

A presenter is visible in the bottom right corner of the slide area.

Now, I will share some future directions. So, group level emotion is actually a fairly new area as we are getting more and more data sets collected. For example, here is one from Sharma and others and then there are other repositories such as there is a happy data set and then there is a group affect data set GAF. So, these are available publicly. You can, you know, download them, train systems and test out group emotions.

Now, we discussed about image level group emotion, but of course, we have discussed when you have time series information that is giving you a lot more information for understanding the emotion of a person or a group of people. So, in this case as well, if you have a video of a group of people, you can analyze the rich dynamics with respect to the group and then add that information to the expression.

The other direction friends is audio video fusion. So, generally in group settings you will have audio information available as well. Now, this audio information would be in the form of either the voices of someone could be speaking in a group of people or it could be some background music. So, that information can actually help us in more robust facial expression recognition right. So, it will be complementary information.

The third is more larger data sets are required. Typically, what happens is when you are collecting a data set where you have a group of people and you are trying to label the perceived emotion of those people, there is a lot of variability which can come. Now, this variability would be in the form of the labels which different labels, labelers could be giving to the same video.

So, person 1 could perceive a video having let us say very high positive perceived emotion. Person 2 could have it as a medium positive perceived emotion right. Then how do you get the consensus right? How do you get the consensus among the labelers? That is important. The other is how do you decide that one labeler is more consistent as compared to other labeler right.

So, if you know a labeler one is more consistent in creating the data set is having more consistent labels then you could perhaps give more weightage to a labeler who is more consistent. And the third is when you are collecting these data sets what about the copyright? Right, you just cannot download videos and use it without proper copyright check.

So, typically in the academic research community we will look for data with creative commons license. So, you can dwell deeper into it and then make sure that the large data set which you are collecting you actually have the right for it. So, friends this brings us to the end of the group emotion part of the automatic facial expression recognition lecture.

Thank you.

