Maastricht University

Investigating the Explainability Power of Transformers in the Video Domain

RESEARCH INTERNSHIP ARTIFICIAL INTELLIGENCE

Author
Christopher DU TOIT

Supervisor
Dr. Mirela POPA

This paper introduces the pure self-attention mechanism to emotion detection in the video domain for the first time. The Timesformer is trained on the high quality dataset CREMA-D, scoring a high accuracy, and then various visual techniques are adapted and applied to the model. Occlusion testing is then performed on the model to quantitatively measure the performance of the visual techniques as-well-as to find which parts of the human face are important to the transformer architecture for accurately identifying the emotion shown.

February 2022

Contents

1 Introduction

Today, Explainable Artificial Intelligence (XAI) is more important than ever before. With a multitude of models being developed every year, more and more insight into how exactly these models work, and make their predictions, is required for the models to be utilized in the real world. In addition, the need for XAI is ethically relevant especially as machine learning models start to take over or assist ethically sensitive professions such as law, medicine or surveillance etc..

A particularly new model architecture is the self-attention-based architecture, namely Transformers, which have been predominantly used in Natural Language Processing [7], but have recently seen great success in Computer Vision, mainly due to their efficiency in computation. The Timesformer [2], is an extension of the Vision Transformer (ViT) [8] that purely uses the self-attention, i.e. no added convolutions [15], [14], mechanism for the classification of videos. With the recent advancement of transformers in Computer Vision, particularly in image classification, there has been an effort to visualize where on the images the transformer *'pays attention'* to to make it's decision. In the video domain, there has been little to no work done at the time of writing this report.

'Pure' transformer models have not been trained for emotion detection in images nor in videos. Therefore, this project is the first to explore the behaviour of the Timesformer on emotion detection. Although there are many available emotion datasets, CREMA-D [3] offers a professional dataset containing good video and sound quality with a clear focus on the actor and a clear background. 6 different emotions make up the dataset; happy, sad, angry, disgust, neutral and fear.

To understand how the self-attention mechanism learns each emotion, various visual techniques designed for image classification will be adapted to the video domain as this has also not been done before. In addition, these visual techniques will be quantitatively measured by occlusion testing. The final part of the project will be to test occlusion of the individual facial parts and then test the model to see which emotions are no longer accurately identified because of such occlusion. With this, novel trends about how the model identifies each of the different emotion types will be analyzed and discussed.

2 Methodology

In this section, the problem statement and research questions are stated. Then, background information on the Timesformer, CREMA-D dataset and visual techniques, namely Grad-CAM, Rollout and TIBAV are given as-well-as details as to how the Timesformer is trained and the performance of the model.

2.1 Problem Statement and Research Questions

With Transformer models being still relatively new in the field of Computer Vision, there is room for exploration and analysis of visual techniques to get a better understanding of the self-attention mechanisms. In particular, the goal of this project is to focus on the detection of emotions in the video domain and to try and find trends in the way the model identifies each of the emotion types. With that being said, the problem statement is:

How does the self-attention mechanism identify emotions?

To answer this, the following research questions shall be addressed.

- What is the performance of the TimeSformer model on detecting emotions when trained and tested on the CREMA-D dataset?
- How can visual techniques such as Grad-CAM be extended to the video domain?
- Can the reasoning behind the prediction outputs of the TimeSformer model be visualized using the extended visual techniques?
- Can occlusion testing be used to quantitatively evaluate the visual technique performances?
- Based on the performed analysis, can we obtain a set of visual patterns for each emotion?

2.2 TimeSformer

The TimeSformer is an extension of the standard Vision Transformer (ViT), where the idea behind ViT was to use the standard Transformer, which is predominantly used in Natural Language Processing (NLP), and apply it to the image domain with as little change as possible. This makes the Timesformer a suitable model for this project as it is as 'pure' a transformer model as possible. The way the ViT works is, instead of inputting a sequence of words into the Transformer, they cut an image up into equal patches and feed these patches as embedded tokens into the model. Then the model computes the query $\mathbf{q}$, key $\mathbf{k}$ and value $\mathbf{v}$ matrices which are used to compute the attention between each patch and every other patch. Therefore, for an image with N by N patches, for each patch there are $N+1$ computations of attention. The $+1$ counts for the classification token which is used to give the output classification of the image. This is repeated for L encoding blocks and A attention heads.

$$\boldsymbol{\alpha}_{(p)}^{(l,a)} = SM\left(\frac{\mathbf{q}_{(p)}^{(l,a)\top}}{\sqrt{D_h}} \cdot \left[\mathbf{k}_{(0)}^{(l,a)} \left\{\mathbf{k}_{(p')}^{(l,a)}\right\}_{p'=1,\ldots,N}\right]\right)$$

Here, α is the attention, $l \in L$ is the block, $a \in A$ is the attention head and p is the patch. SM represents a SoftMax computation. Then, a weighted sum of the value matrices and the self-attention coefficients $\boldsymbol{\alpha}_{(p)}^{(l,a)}$ is computed.

$$\mathbf{s}_{(p)}^{(l,a)} = \boldsymbol{\alpha}_{(p),(0)}^{(l,a)} \mathbf{v}_{(0)}^{(l,a)} + \sum_{p'=1}^{N} \boldsymbol{\alpha}_{(p),(p')}^{(l,a)} \mathbf{v}_{(p')}^{(l,a)}$$

Timesformer is extends the use of the ViT from images to videos. Moreover, instead of computing the attention between patches in one frame (image, in the case of ViT), attention is computed between every patch as-well-as in every frames. Therefore, joint Space-Time attention computes attention for each patch $NF+1$ times, where F is the total number of patches. Below is the full Time-Space attention formula.

$$\boldsymbol{\alpha}_{(p,t)}^{(l,a)} = SM\left(\frac{\mathbf{q}_{(p,t)}^{(l,a)\top}}{\sqrt{D_h}} \cdot \left[\mathbf{k}_{(0,0)}^{(l,a)} \left\{\mathbf{k}_{(p',t')}^{(l,a)}\right\}_{\substack{p'=1,\ldots,N \\ t'=1,\ldots,F}}\right]\right)$$

Where t is the frame. The corresponding weighted sum of the value matrix and $\boldsymbol{\alpha}_{(p,t)}^{(l,a)}$ is,

$$\mathbf{s}_{(p,t)}^{(l,a)} = \boldsymbol{\alpha}_{(p,t),(0,0)}^{(l,a)} \mathbf{v}_{(0,0)}^{(l,a)} + \sum_{p'=1}^{N} \sum_{t'=1}^{F} \boldsymbol{\alpha}_{(p,t),(p',t')}^{(l,a)} \mathbf{v}_{(p',t')}^{(l,a)}.$$

Here the attention is computed between every patch and every frame. However, the creators of the Timesformer stated that the full Space-Time attention method is not computational efficient and results in an under-performing model when compared to the disjoint Space-Time attention model. For this project, the disjoint Space-Time method is used, where for each patch, attention is computed between that patch and every other patch in that frame and also between that patch and only that same patch in other frames. This results in a computation of only $N+F+1$ for each patch which is significantly less. In practice, the model computes that temporal attention, then updates the key, query and value matrices before computing the spacial attention implying that the output spacial attention is all that is needed for the visualization as it inherits the temporal attention as well.

2.2.1 CREMA D

CREMA-D is a data set containing 7,442 original short videos of about 2-3 seconds long. In each of the videos, there is one of 91 different actors expressing an emotion that is either happy, sad, angry, fear, disgust or neutral. This type of data set is a non-spontaneous data set meaning that the emotions shown are not genuine emotions but are emotions that the actors are told to express. The 91 actors consists of 48 males, 43 females between the ages of 20 and 74 with various races and ethnicities. The videos were then labelled through a crowd-sourced voting method and 95% of the videos have more than 7 ratings.

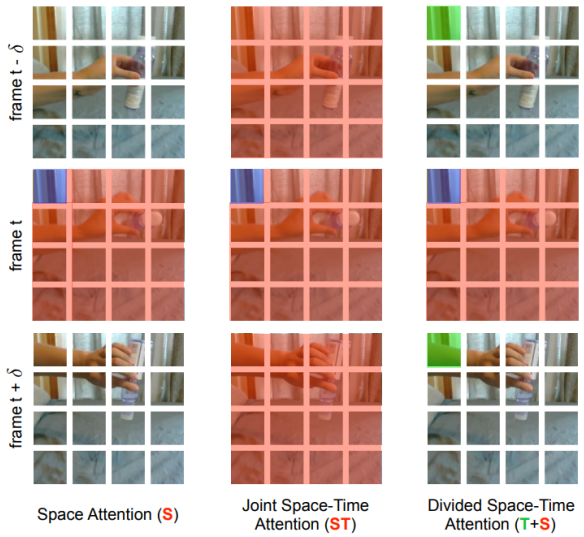


Figure 1: The blue patch is the patch in question and attention is computed between this patch and the other red and green patches. In the first column, a space only attention is computed where the attention is only computed between the patches in the spacial domain. The second column represents the full, computationally heavy joint space-time attention where the attention is computed between the patches in both the spacial and temporal domain. Lastly, the 3 column represents the divided space-time attention mechanism, which is much more efficient. Here, the spacial attention is computed as well as attention between the blue patch and the corresponding blue patch in the other frames, namely, the green patches. This image is taken from [2].

2.2.2 Model Performance

First, 8 evenly spaced out frames were extracted from the videos and then fed through the FaceNet [12], model which cropped the frames to only contain the faces. This preprocessing technique helped reduce the workload of the model during training which enabled training on a personal GPU. These frames were then used for training the non-pretrained TimeSformer. Ideally, it would be have been much better to take a pretrained Timesformer model which has been pretrained on some sort of emotional data set to be able to utilize the power of transfer learning but at the time of the project, there were no pretrained models of such available. The data set was split randomly into 90% training, 5% testing and 5% validation. This produced an 81% model accuracy. Initially there was an idea to do some more data preprocessing like key frame extraction but for the purposes of this project, 81% accuracy is sufficient enough to be able to study different visualization techniques on it.

2.3 Explainable AI and Visual Techniques

Explainable AI (XAI) is quite a vast topic with many different definitions of what exactly XAI is. In this project, XAI will be defined as a heatmap generated from a model that accurately shows where the model looks on the person's face to decide the emotion shown.

Visualizations of Convolutional Neural Networks (CNNs) [10] has been quite successfully researched with Grad-CAM (Gradient-weighted Class Activation Mapping) [13], Grad-CAM Plus Plus [4] and EigenCAM [9] to name a few. However, with the recent introduction of transformers in Computer Vision, there is still progress to be made for visualizations, especially in the video domain. The difficulty lies in how exactly should one extract the attention, which contributes to the classification token, from each attention block and attention head and how do we then combine these attention maps together. For this project, 3 main visualization techniques are studied, adapted and utilized. These are explained below.

2.3.1 Grad-CAM

Grad-CAM was primarily designed for CNN based models but is to be used for the project as a base-line model to compare with. In brief, the Grad-CAM method uses the gradients of the score for a class c with respect to feature map activations of the convolutional layer (usually the last layer). The gradients are averaged over the width and height dimensions to obtain the neuron importance weights. These weights are then applied to the forward activation maps to produce a course heatmap of the activations. Grad-CAM is consequently a class specific method.

Grad-CAM Plus Plus is very similar to Grad-CAM but uses second order derivatives. Lastly Eigen-CAM is also similar to Grad-CAM but instead only takes the first principle component of the 2D activations and therefore has no class discrimination [9].

2.3.2 Rollout

Rollout [1] was one of the first visualization methods designed purely for transformers, but for text as opposed to images. Due to the nature of the self-attention mechanism, the attention weights alone are unreliable as explanation probes. The standard Rollout method uses attention maps only and then aggregates the attention maps by recursively multiplying them together resulting in a matrix called the 'attention rollout'. To compensate for the residual connections, an identity matrix is added to each of the attention maps before aggregation. Lastly, there are multiple attention heads and hence multiple attention rollout matrices. The paper which introduces this method [1], states that one can use the maximum, minimum or average of the attention heads and the choice seems to depend on the matter at hand. Jacob Gildenblat made a post online about using the Rollout method for ViT and showed that taking the average about the attention heads results in a noisy heatmap and that one should take maximum or minimum attention head instead. It should be noted that this is not a class specific method due to the lack of use of gradients.

In addition, Jacob Gildenblat takes some inspiration from TIBAV [6], section 2.3.3, and produces a new class specific version of the Rollout method, which we will call Rollout-Grad. Rollout-Grad first multiplies the attention map with the attention gradient before aggregating it with the attention rollout.

2.3.3 TIBAV

TIBAV [6], [5] stands for the paper "Transformer Interpretability Beyond Attention Visualization". TIBAV offers two class specific methods. The first is their initial technique designed for ViT and incorporates Layer-wise Relevance Propagation (LRP). However, this technique is very difficult to apply to other transformer models such as the Timesformer. Fortunately, they later released a much simpler method which also provided great results by using 2 rules and this method will be the only TIBAV method investigated.

$$\bar{\boldsymbol{\alpha}} = \mathbb{E}_h((\Delta\boldsymbol{\alpha} \odot \boldsymbol{\alpha})^+)$$
$$\mathbf{R} = \mathbf{R} + \bar{\boldsymbol{\alpha}} \cdot \mathbf{R}$$

The first rule shows that a Hadamard product ($\odot$) between the attention map and the gradient $\Delta\boldsymbol{\alpha} = \frac{\delta y_t}{\delta \boldsymbol{\alpha}}$, where y_t is the models output class. Then, all negative values from this multiplication are set to 0 to remove any negative contributions and finally an average across the h attention heads is computed. In the second rule, $\mathbf{R}$ is the relevancy, which is initialized as the identity matrix, and is updated by adding the matrix multiplication of the relevancy matrix by $\boldsymbol{\alpha}$ to itself.

3 Experiments

In this section, the experimental procedures explored will be stated and reasoned as well as goals for outputs of such experiments. The results of these experiments can be found in the following section.

3.1 Adapting and testing visual techniques

Each of the visual techniques described in section 2.3 will be adapted to the video domain and examples of each of the visualizations will be produced. This will give some insight into the behaviour of the techniques.

In addition, the procedures behind how each of these visual types were adapted for the video domain will be explained.

3.2 Visual technique occlusion testing

In attempt to evaluate how well a visual technique accurately encapsulates the important patches of the image, occlusion testing will be applied to the visualization techniques Grad-CAM (which will act as a base case), Rollout, Rollout-Grad and TIBAV. Occlusion testing works in 2 ways; the first is to occlude areas of the frames that the visualization technique has claimed to be important to the classification of the video. If the model is then not able to identify the emotion of the video with this occlusion then it could imply that the visualization technique does a good job of identifying the important parts of the frames.

The second part to this is to occlude the parts of the frames where the visualization technique claims are not important for the classification of the video. If the video is not able to identify the emotion presented in the video then it could imply that the visualization technique is not accurately extracting the correct activations or attention from the model.

Therefore, to explain how the experiments will go, we first must define what most important patches and least important patches mean. The most important patch is a patch that a particular visual technique gives high heat map values to and consequently, the least important patches and the heat map values with the lowest values. Similarly, because there are multiple frames, there will also be a most important frame and least important frame which has the highest sum of heat map values and lowest sum of heat map values respectively. Therefore, the experiment will go as follows:

Experiment 1: The top 0, 10, 20, 30, 40, 50 percent of the most important patches will be occluded from the test data set. For each occlusion percentage, the accuracy of the test data set will be measured. Then the experiments will be repeated but with the top 0, 1, 2, 3, 4 frames being occluded and again the test accuracy will be recorded. The expectation here is that there will be a steep drop in accuracy as more of the important patches and frames are occluded.

Experiment 2: Almost identical to experiment 1 but instead of the most important patches/frames, it will be the least important. The expectation is that there will be a much less steep drop in accuracy as more of the non-important patches/frames are occluded.

The best performing visual technique will be the model that exceeds in both experiments and hence the results should include a metric which takes the difference between the performance of these 2 experiments.

3.3 Emotion trends

The best performing technique from section 3.2 will be then used to find general trends across the emotion types. The experiments include storing the heat maps for all the correctly predicted videos from the test data set and then averaging all the heat maps of the same emotion together to find a general trend. In addition, a line graph of the frame activity will be plotted for each emotion.

3.4 Facial part occlusion testing

The final experiment is to try and see which key facial parts are instrumental in detecting a particular emotion. To do this, specific facial parts and combinations of these facial parts will be occluded. The facial parts are both eyes together, the nose and the mouth, extending all the way out to the sides almost as if the person is wearing a mask. For each experiment, an occlusion matrix will be produced.

4 Results

In this section, the results of the experiments described in the previous section are shown and analyzed.

4.1 Adapting and testing visual techniques

4.1.1 Grad-CAM

Since each of the Grad-CAM model only considers the final layer (or block in the case of transformer architecture), there was virtually no adaptation as the output of the visualization was already of the correct shape. Examples of each of the Grad-CAM visualizations can be found in figures 2 and 3. In these examples, one can see that Grad-CAM and Grad-CAM Plus Plus are very noisy visuals whereas Eigen-CAM is more specific and accurate. This clearly shows that the Grad-CAM visual methods are not so good for self-attention models as the attention propagating through all the layers is important for the final visualization. More examples of the other emotions can be found in the appendix in section 7.1.1.

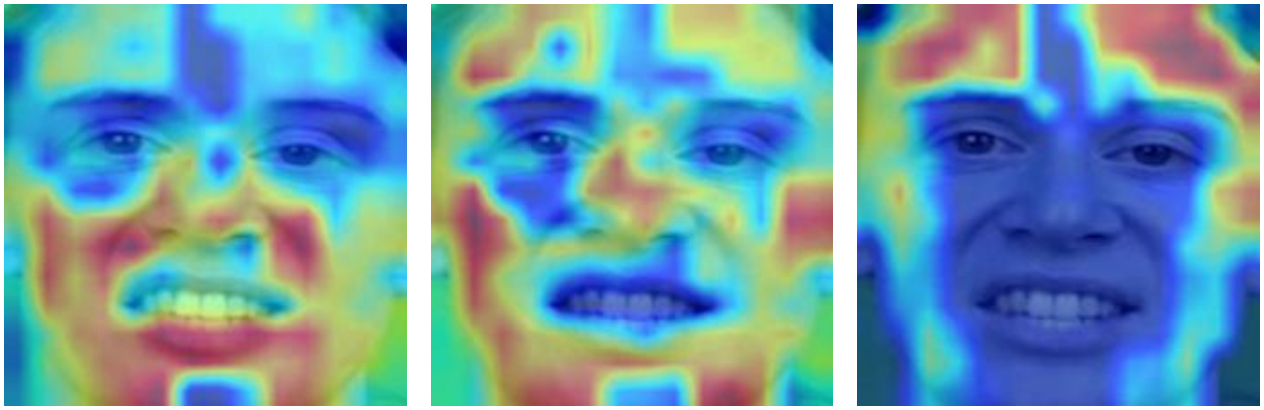


Figure 2: Visualizations for a frame of the anger emotion, and from left to right it is Grad-CAM, Grad-CAM Plus Plus and Eigen-CAM.

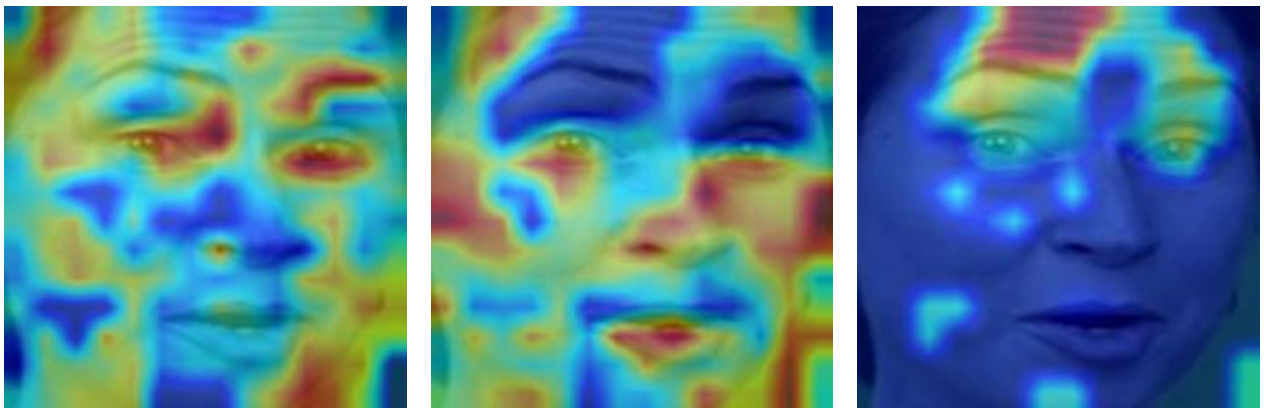


Figure 3: Visualizations for a frame of the happy emotion, and from left to right it is Grad-CAM, Grad-CAM Plus Plus and Eigen-CAM.

4.1.2 Rollout

For the standard Rollout method, the minimum attention head is used because the other 2, namely the maximum or average, produced noisy outputs. Examples of each Rollout method can be found in figures 4 and 5. One can immediately notice that these visual techniques are far more specific than the Grad-CAM visualizations, and in these particular frames, the standard Rollout seems to produce 'better' visuals than Rollout-Grad. The implementation of both of these visualizations proved to also be simple since they were designed for transformer models. The only change needed was to make sure that the models extracted the

correct attention maps as the Timesformer contains both temporal attention maps and spacial attention maps. More examples of the other emotions can be found in the appendix in section 7.1.2.

Figure 4: Visualizations for a frame of the anger emotion, with Rollout on the left and Rollout-Grad on the right.

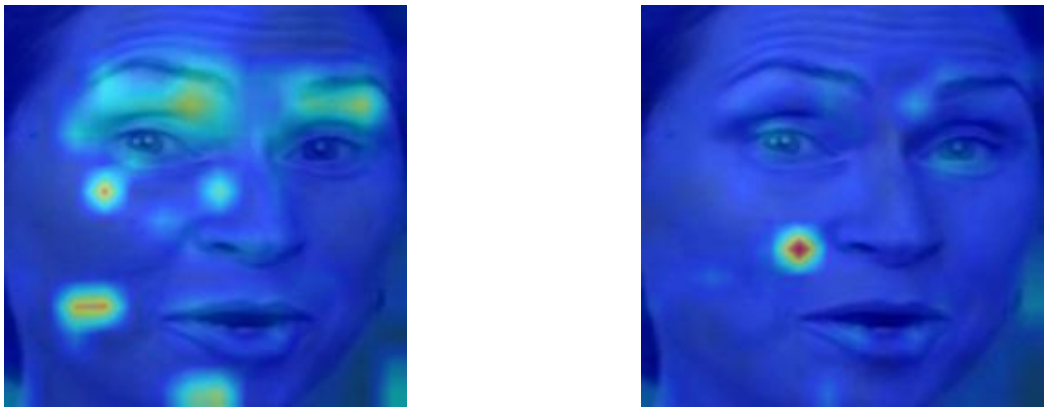


Figure 5: Visualizations for a frame of the happy emotion, with Rollout on the left and Rollout-Grad on the right.

4.1.3 TIBAV

Figure 6 shows the TIBAV visuals on the example frames. It is clear that TIBAV produces the most complete visualization and this is especially noticeable in the happy emotion where it shows that the eyebrows are of interest as-well-as the top lip and edges of the mouth. The adaption of the TIBAV model took the most work as it meant editing the Timesformer model such that it could save the correct attention maps and also the attention map gradients. More examples of the other emotions can be found in the appendix in section 7.1.3.

Figure 6: TIBAV visuals of the emotions anger on the left and happy on the right.

4.2 Visual type occlusion testing

Figures 7 and 8 show examples of the occlusions for each of the visual types where only 20% of the most important patches are occluded and similarly, figures 9 and 10 show for the 20% of the least most important patches. In figures 7 and 8 it is clear that there is some focus on the mouth, eyes and area between the eyes for the anger emotion, with an emphasis on the mouth. For the happiness, there is more of an emphasis on the eyes than the mouth but the mouth is also included. Whereas for figures 9 and 10, the occlusion is in the parts of the image which are clearly not important (except in the case of Grad-CAM). This is in line with the knowledge from affective computing which states that anger is, *'characterized by inward lowering motion of the eyebrows coupled with compaction of the mouth'* [11]. For all emotion examples, please refer to the appendix section 7.2.

Figure 7: Examples of 20% occlusions of the most important patches for the anger emotion for Grad-CAM, Rollout, Rollout-Grad and TIBAV from left to right.

Figure 8: Examples of 20% occlusions of the most important patches for the happy emotion for Grad-CAM, Rollout, Rollout-Grad and TIBAV from left to right.

Figure 9: Examples of 20% occlusions of the least important patches for the anger emotion for Grad-CAM, Rollout, Rollout-Grad and TIBAV from left to right.

Figure 10: Examples of 20% occlusions of the least important patches for the happy emotion for Grad-CAM, Rollout, Rollout-Grad and TIBAV from left to right.

The figures in figure 11 show the testing accuracy of the model on the occluded data. Each plot represents a different percentage of spacial occlusion and one can see that as there is more spacial occlusion, the max occlusion (A.K.A. the occlusion of the most important patches) accuracy decreases drastically as expected and the min occlusion (A.K.A. the occlusion of the least important patches) accuracy only decreases marginally. Moreover, the accuracy drops as number of frames occluded are increased but the difference is not as significant, showing that there is less of a dependence on the temporal aspect of the model.

It should be noted that Grad-CAM performs significantly worse than the other models. The general trend over the subplots in figure 11 is that as there is more occlusion, there should be a larger difference between the 2 lines but for Grad-CAM, these lines actually stay very close to each other. This implies that the Grad-CAM visualization is not accurately extracting the correct attention maps from the model. On the other hand, TIBAV performs significantly better. This is because it manages to retain a high accuracy of about 63% when the spacial occlusion for the least important patches is at 50% while scoring a very low accuracy of about 33% for 50% occlusion of the most important patches. Hence, this implies that the TIBAV

accurately extracts the self-attention behaviour of the Timesformer model.

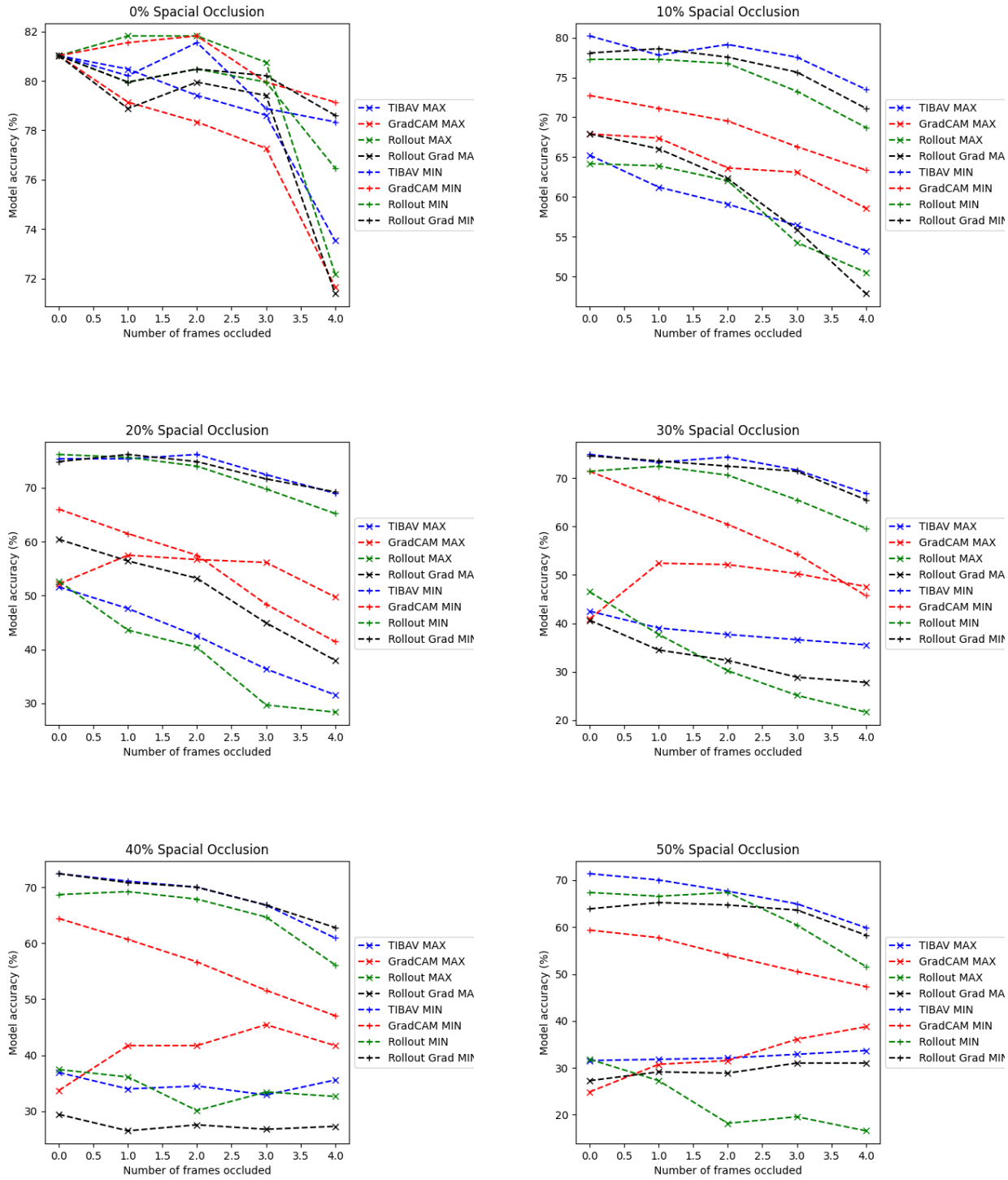


Figure 11: The figures above show the accuracy of the most/least important occlusions on the test accuracy. Each plot is an increase in 10% spacial occlusion and in each plot, the y-axis represents the testing accuracy and the x-axis represents the number of frames occluded.

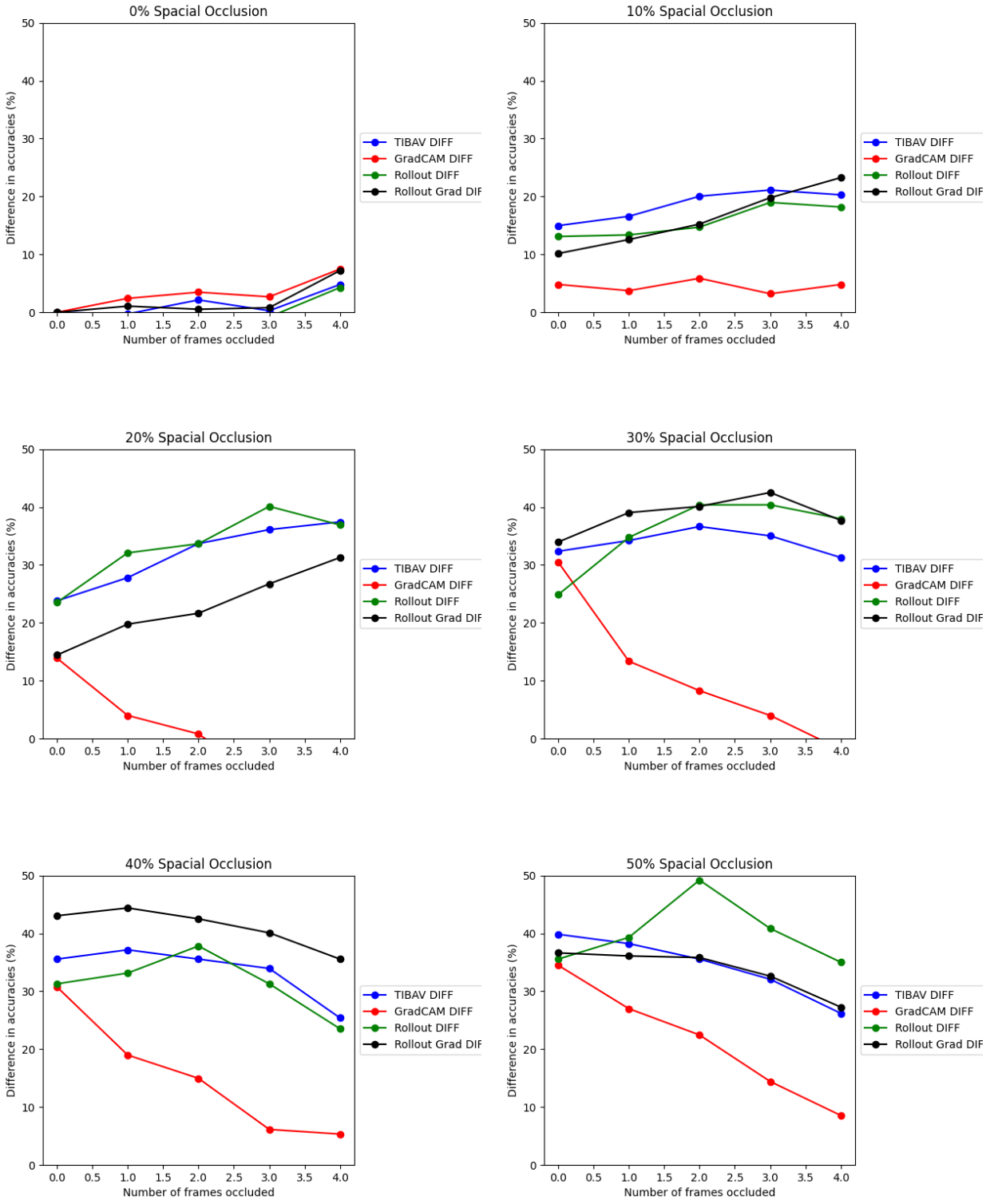


Figure 12: The figures above show the difference in accuracy of the most/least important occlusions on the test accuracy. Each subplot is an increase in 10% spacial occlusion and in each plot, the y-axis represents the min testing result minus the max testing result and the x-axis represents the number of frames occluded.

The figures in figure 12 show the accuracy result of the least important occluded frames minus the accuracy result of the most occluded frames. Therefore, the highest line is the best performing visualization technique. These plots help us to understand and see that Grad-CAM is clearly inferior to the other 3 visual techniques. It also shows that the blue line, representing TIBAV, is on average the highest performing visual technique as it outperforms the other visual techniques more than it is outperformed by those techniques.

4.3 Emotion trends

In this section, the TIBAV model is used to produce the heat maps shown in figure 13, as-well-as the line plot 14. The heat maps represents the 14×14 patches that are fed into the Timesformer over the 8 frames, and one can see that there are basic facial elements displayed. For instance, in the happy heat maps, one can almost see a smile and 2 eyes. The heat maps are to scale and this shows that there is a large amount of intensity around the mouth of the anger emotion. Heat maps for the remaining emotion types can be found in the appendix in section 7.3. In figure 14, it is clear that happiness and anger highest average of attention intensity but there is no real trend over the 8 frames.

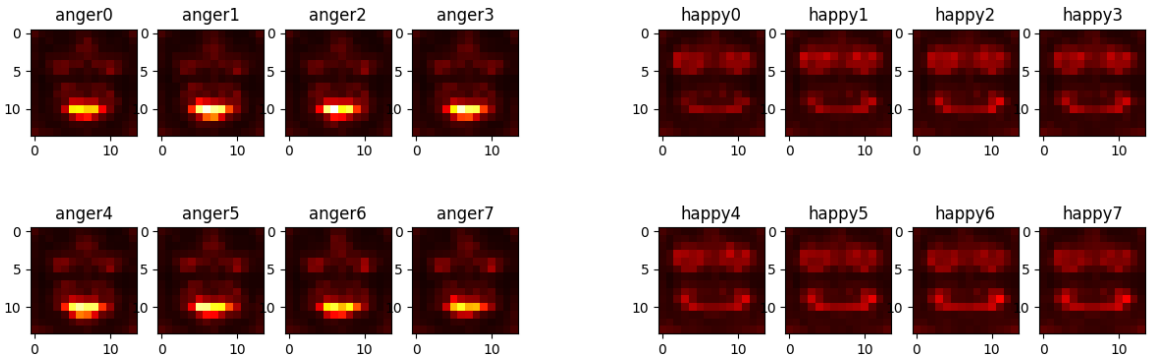


Figure 13: General trends heat maps from TIBAV with anger on the left and happy on the right.

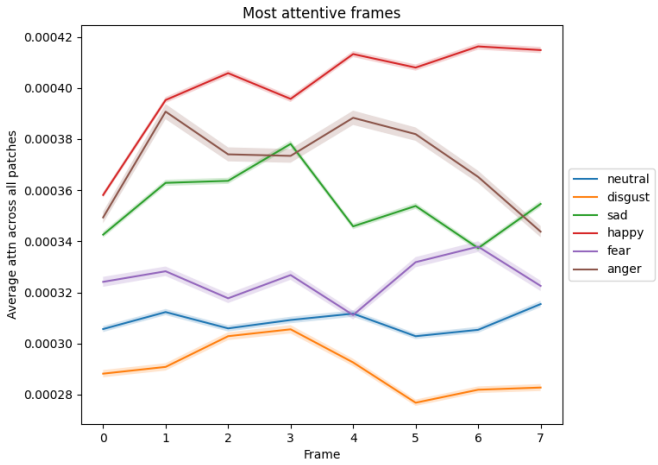


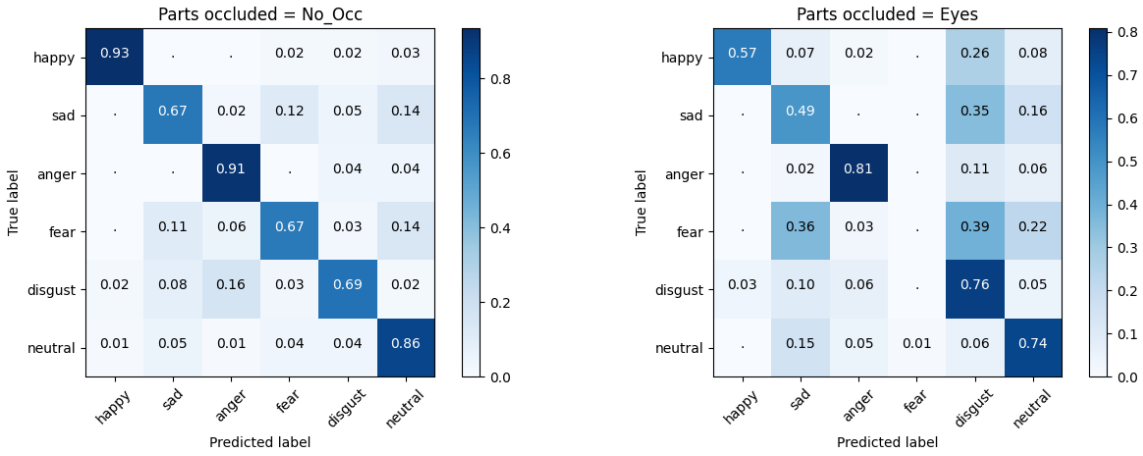
Figure 14: The plot shows the average intensity of emotion per frame with the error bars.

4.4 Facial part occlusion testing

Figure 15 provides examples of the facial part occlusion. It was possible to do this easily due to the fact that the frames were cropped to the face via FaceNet and the actors were always facing towards the camera, therefore implying that the general area of each of the facial parts was mostly the same for each of the videos. The rest of the examples can be found in the appendix in section 7.4.

The figures in figure 16, are the confusion matrices corresponding to the type of occlusion implemented. One can see that as soon as the eyes have been occluded, the model is no longer able to detect the fear emotion and there has been more than 10% drops in accuracy for happiness and sadness. Disgust is the only emotion to receive a better accuracy. When the mouth is occluded, one should note that fear and neutral do not significantly change whereas anger, disgust and happiness do see a change with anger seeing the biggest drop. This makes sense due to the earlier observations in the heat maps, figure 13. Occluding the nose however, results no significant changes to accuracy except for the sad emotion. In the remaining graphs, the trend is that the more occlusion there is, the more the model simply predicts any emotion to be neutral. Some interesting observations are that anger and disgust maintain a high accuracy when the eyes and nose are occluded. In addition, fear has no significant loss in accuracy when the nose and mouth are occluded and lastly, when all 3 facial parts are occlude, no emotion does well except for neutral of course.

Figure 15: Examples of facial parts that have been occluded.



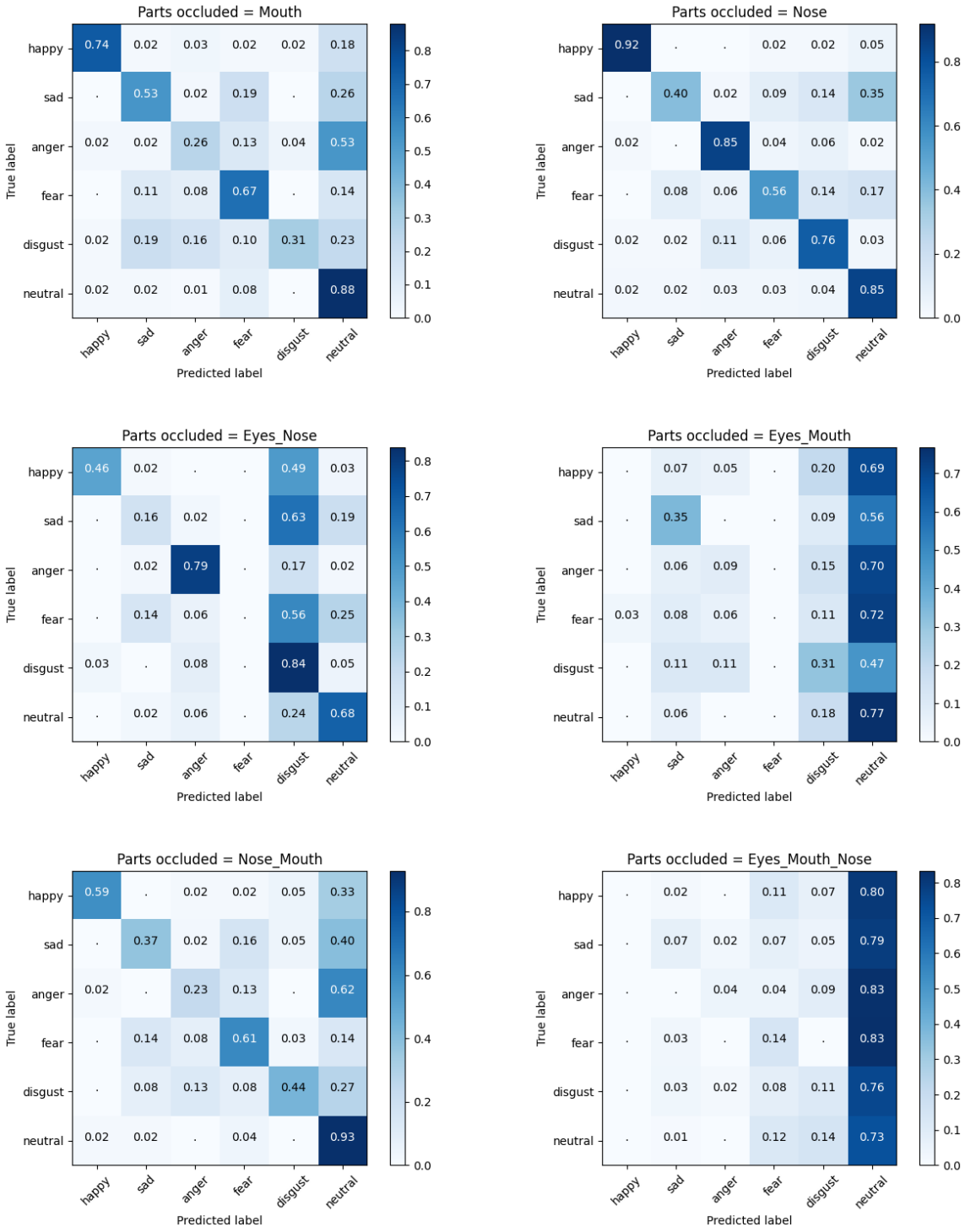


Figure 16: Confusion matrices where the titles show which facial parts have been occluded.

5 Discussion and further research

Adapting each of the visual techniques proved to be a simple task and the reason this could be is that the Timesformer model was trained using the disjoint time-space attention mechanism. Moreover, as explained

in section 2.2, this meant that only the spacial attention was needed for the visualizations which is similar to computing visualizations of each frame as separate images. However, in the case of the full space-time attention, it would have been more difficult to adapt the visual techniques.

It is clear that Grad-CAM is not suitable for the Timesformer but the other 3 visual techniques perform on a fairly similar level to each other, especially TIBAV and Rollout-Grad due to their similarities in architecture. When testing the performance of the visualization techniques via occlusion testing, it was noted that the occlusion of the frames had much less of an impact than the spacial occlusion. This is most likely due to the fact that the model was trained on disjoint Space-Time method which had much less temporal influence on the classification of the video as opposed to the full Space-Time method.

It can be agreed that the model does look at key features of the face to identify the emotions because in the heat maps, one can identify a face and the parts of the face. Moreover, when the specific facial parts were occluded, one could notice that the detection of a particular emotion was either not affected at all or completely affected. For example, the case where the accuracy of prediction a fear emotion was not significantly affected when both the nose and mouth were occluded and vise versa, fear not being detected at all when the eyes were occluded implying that the model very much looks at the eyes to identify fear.

Based on the experiments, a summary of what the model looks at for each of the emotions are:

- Happy: The eyes and mouth.
- Anger: The mouth.
- Sad: The eyes and nose.
- Fear: The eyes.
- Disgust: The mouth.
- Neutral: The whole face - not well defined.

It is interesting to see that anger and disgust are both only identified by the mouth as these two emotions are arguably quite similar in characteristics. These results are in line with the average heat maps produced by the TIBAV visualizations. For instance, in the left subplot of figure 13, it is clear that the focus is of the mouth and in the right subplot, the focus is on the eyes and mouth. TIBAV also agrees with the above observations for the emotions disgust and neutral, figures 31, 32. However, in the case of fear, the subplot on the right of figure 32 shows that the focus is of the eyes and mouth which is in contrary to the above observations that states that fear only looks at the eyes. And similarly, for the sad emotion, TIBAV seems to highlight the part in between the eyes and the mouth but the above states that the eyes and nose are the important parts.

Finally, some further research suggestions would be to fine-tune this trained Timesformer model on a spontaneous emotion data set, where the emotions displayed are genuine and not acted. This may not perform as well as this current model but it could bring some interesting insights into how different and harder genuine emotions are to detect.

6 Conclusion

In this project, the Timesformer was trained on the emotion dataset CREMA-D with the disjoint space-time attention mechanism and produced an accuracy of 81%. Different visual techniques, namely, Grad-CAM, Rollout and TIBAV were adapted to the video domain and then applied to the model. It was noted that the disjoint space-time attention mechanism simplified both the adaptation of the visual techniques as-well-as the model training but still managed to produce great results. This implied that the Timesformer did not need to depend too much on the spacial attention for accurate emotion detection.

Each of these visual techniques were then measured both qualitatively and quantitatively, where the quantitative analysis was done by occlusion testing. Grad-CAM performed the worst overall and TIBAV performed the best and this was largely because TIBAV considered all the attention maps and gradient maps throughout the transformer architecture whereas Grad-CAM only considered the final layer. Specific facial

parts were then occluded to try and find trends in each of the emotions and a clear pattern for each of the emotions was discovered. The general trends produced by TIBAV agreed with these patterns for all emotions except for the fear and sad emotions. Lastly, some further research for emotion detection with transformers was suggested.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers, 2020.

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?, 2021.

[3] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset, 2014.

[4] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar 2018.

[5] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers, 2021.

[6] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization, 2021.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[9] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. *2020 International Joint Conference on Neural Networks (IJCNN)*, Jul 2020.

[10] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks, 2015.

[11] Rosalind W. Picard. Affective computing. *ISBN 978-0-262-16170-1*, 1997.

[12] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015.

[13] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct 2019.

[14] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, 2021.

[15] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers, 2021.

7 Appendices

7.1 Visual techniques examples

7.1.1 Grad-CAM

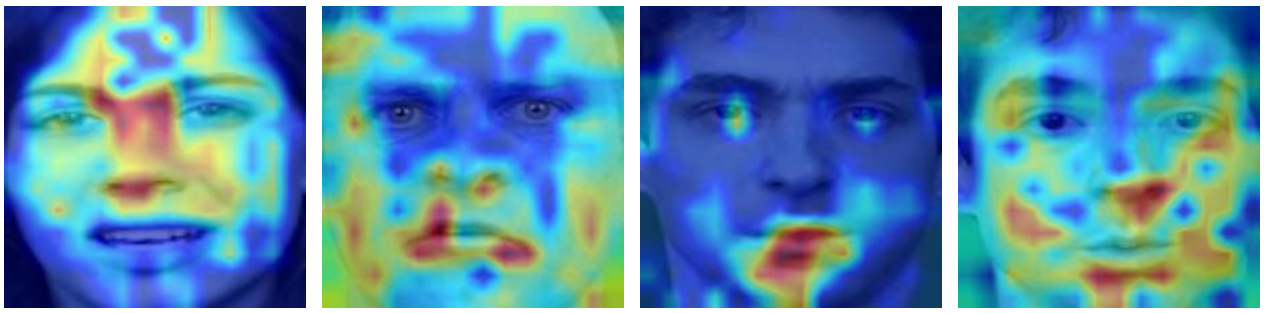


Figure 17: Examples of Grad-CAM visualizations on sad, dear, disgust and neutral from left to right.

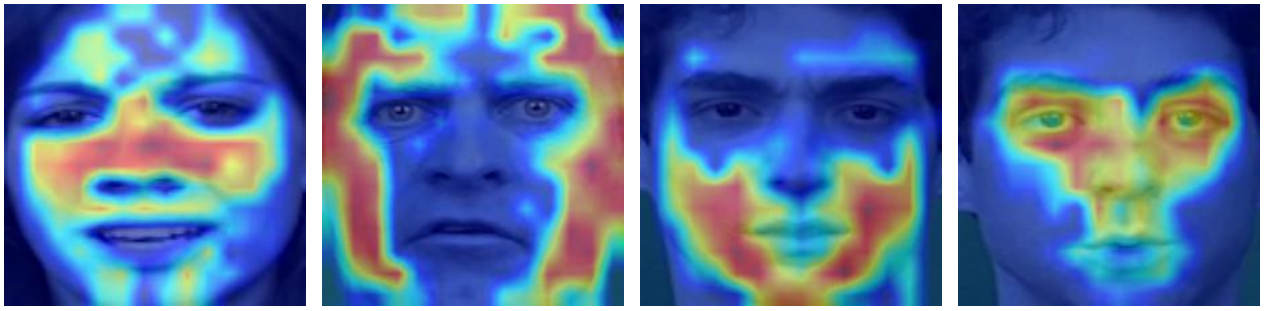


Figure 18: Examples of Eigen-CAM visualizations on sad, dear, disgust and neutral from left to right.

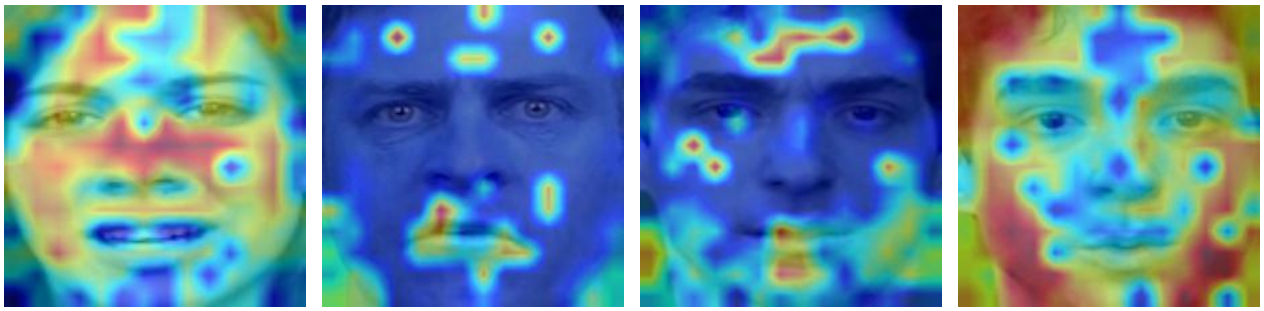


Figure 19: Examples of Grad-CAM Plus Plus visualizations on sad, dear, disgust and neutral from left to right.

7.1.2 Rollout

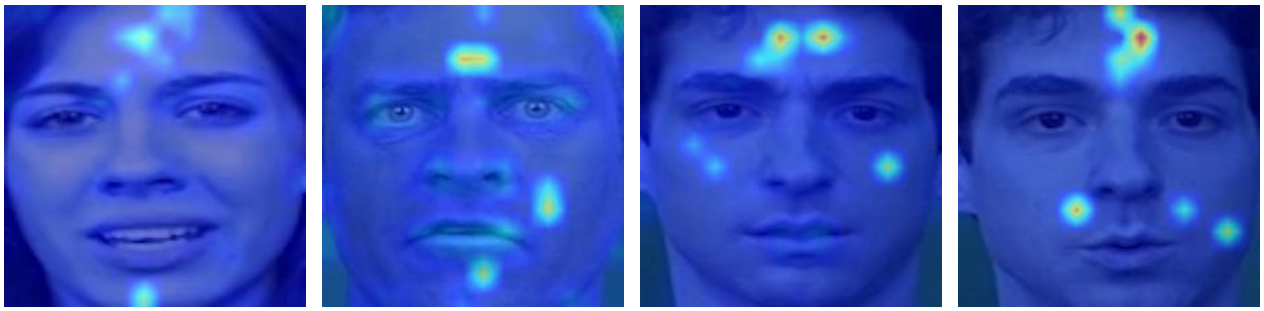


Figure 20: Examples of Rollout visualizations on sad, dear, disgust and neutral from left to right.

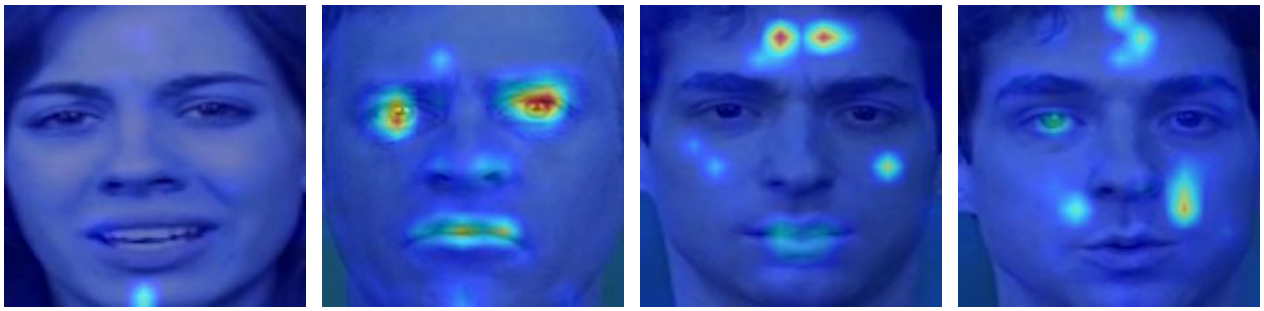


Figure 21: Examples of Rollout-Grad visualizations on sad, dear, disgust and neutral from left to right.

7.1.3 TIBAV

Figure 22: Examples of TIBAV visualizations on sad, dear, disgust and neutral from left to right.

7.2 Occlusion examples

7.2.1 Top 20% most important patches occluded

Figure 23: Examples of 20% occlusions of the most important patches for the sad emotion for Grad-CAM, Rollout, Rollout-Grad and TIBAV from left to right.

Figure 24: Examples of 20% occlusions of the most important patches for the fear emotion for Grad-CAM, Rollout, Rollout-Grad and TIBAV from left to right.

Figure 25: Examples of 20% occlusions of the most important patches for the disgust emotion for Grad-CAM, Rollout, Rollout-Grad and TIBAV from left to right.

Figure 26: Examples of 20% occlusions of the most important patches for the neutral emotion for Grad-CAM, Rollout, Rollout-Grad and TIBAV from left to right.

7.2.2 Top 20% least important patches occluded

Figure 27: Examples of 20% occlusions of the least important patches for the sad emotion for Grad-CAM, Rollout, Rollout-Grad and TIBAV from left to right.

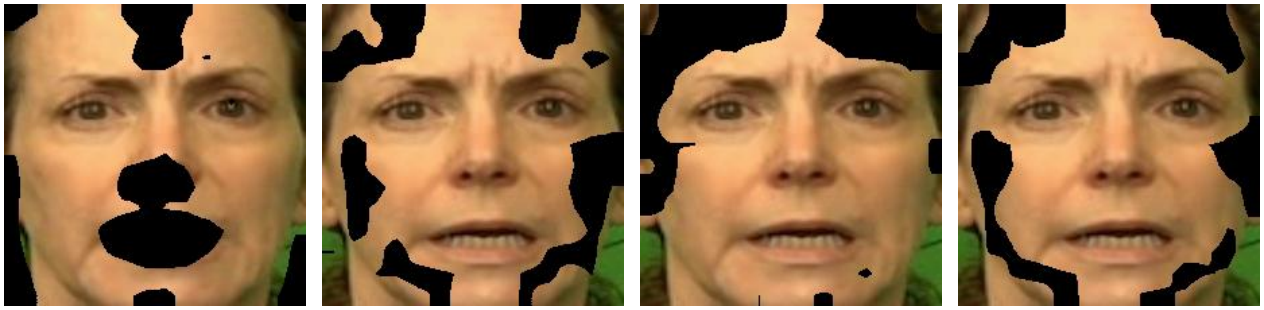


Figure 28: Examples of 20% occlusions of the least important patches for the fear emotion for Grad-CAM, Rollout, Rollout-Grad and TIBAV from left to right.

Figure 29: Examples of 20% occlusions of the least important patches for the disgust emotion for Grad-CAM, Rollout, Rollout-Grad and TIBAV from left to right.

Figure 30: Examples of 20% occlusions of the least important patches for the neutral emotion for Grad-CAM, Rollout, Rollout-Grad and TIBAV from left to right.

7.3 General trends heat maps

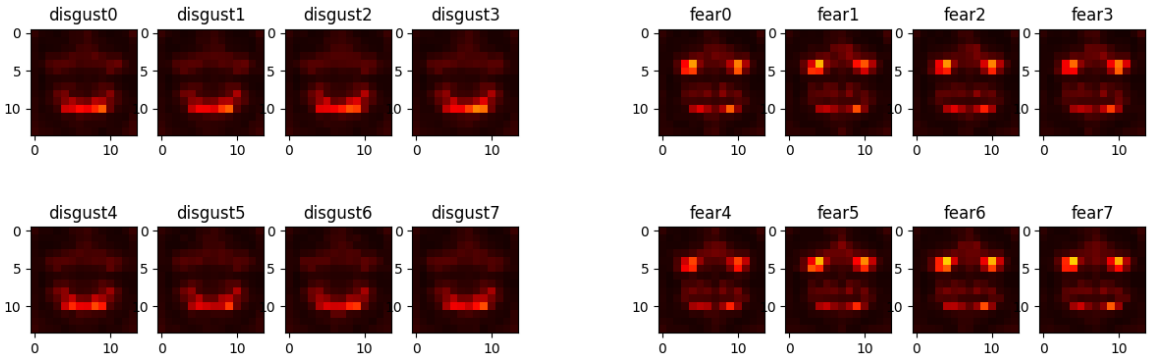


Figure 31: General trends heat maps with disgust on the left and fear on the right.

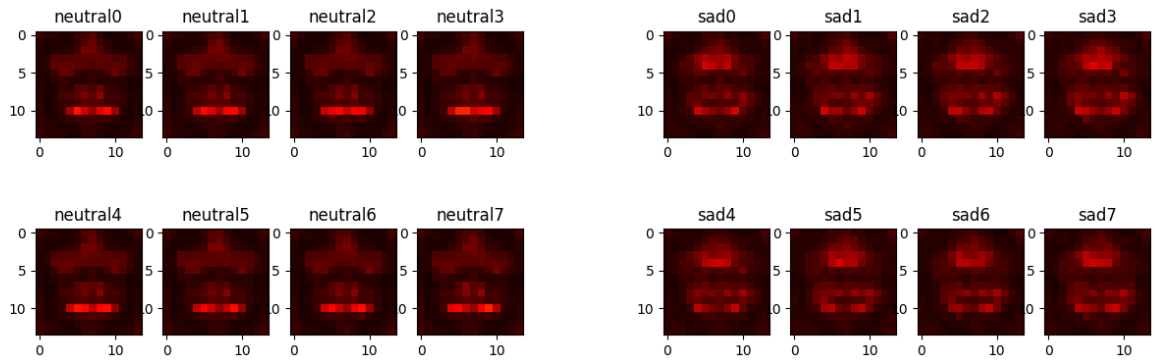


Figure 32: General trends heat maps with neutral on the left and sad on the right.

7.4 Facial part occlusion testing

Figure 33: Example of video with parts of the face occluded. From left to right we have *Eyes and Mouth*, *Eyes and Nose*, *Nose and Mouth* and *Eyes, Nose and Mouth* occlusion.