Building an Artificial Intelligence System to Detect Human Faces in Video & Predict Video Memorability

Kevin Callan #17213076
School of Computing, Dublin City University, Ireland
Email: kevin.callan@mail.dcu.ie

Abstract—**With more of the world's internet traffic comprised of video than ever before, developing the ability to understand and predict video memorability has important applications in fields such as education, advertising, rehabilitation and engineering. The MediaEval data set consists of 10,000 short soundless videos extracted from raw footage. They come with a set of pre-extracted features, such as: HoG descriptors, LBP, SIFT, Color Histogram, Fc7 layer from Inception and C3D features. The data set includes two separate scores for calculating memorability. One score is used as a proxy for shorter term memorability (24 hours after being exposed), while the other is for longer term memorability (72 hours after exposure). The memorability task involves predicting this score and results are based on using Spearman rank correlation coefficient. The aim of this work is to provide an expansion to the the MediEval Media Memorability task by first understanding the data set's existing features and then subsequently engineering and incorporating facial recognition features in an attempt to improve the prediction of video memorability. Finally, this work will model the existing and newly engineered features to provide new insight into the space.**

I. Introduction

A. *MediaEval Background*

MediaEval is a benchmarking initiative, which aims to evaluate newly developed algorithms and approaches to understand human focused and social multimedia. The initiative attracts research teams interested in areas of multimedia analysis such as speech recognition, multimedia content analysis, audio analysis and social networks. In 2018, participants from nine universities and research teams completed and submitted results for the inaugural year of the MediaEval Memorability task [5].

B. *Memorability Task Background*

The MediaEval memorability benchmarking task is concerned with understanding what makes a video more or less memorable, ultimately with the aim of creating an automatic machine learning architecture for calculating a probability score associated with a video's memorability. Effective prediction of memorability will advance our understanding of multimedia content "by putting human cognition and perception at the centre of the multimedia analysis." [5].

It contains two prediction sub-tasks that are analogous to the short and long term memory mechanisms of the human brain. By dividing into two sub tasks, the aim is to achieve a greater understanding of these processes.

- Short-term Memorability: involves predicting a 'short-term' (hours after exposure) memorability score for a given short video.
- Long-term Memorability: involves predicting a 'long-term' (days after exposure) memorability score for a given short video [5].

C. *What Type of Media and the Differences Between Video and Images*

In order to understand the approaches and assumptions discussed in this paper it is helpful to first define the types of media that are being analysed.

Videos are sequences of moving frames, where a temporal relationship exists between the frames [9].

This particular data set is comprised of 10,000 videos, each seven seconds in length. These will have to be processed in order to develop new features. However this presents computational problems due to the intensive nature of calculations over an entire video. Specifically, processing this many raw videos would require large random access memory (RAM) and storage resources.

More simplistically, videos can be defined as a moving sequence of images. The more simplistic definition neglects the fact that relationships exist between the sequences of images.

The solution that this work employs extracts one image for each second in the video and conduct any feature processing on the specific image, then performs operations on those 7 images to calculate features for each video. While this reduces the computational resources required to process raw video, it also presents a potential loss of signal that could help predict memorability. This potential loss arises from not utilising temporal aspects of the video and should be noted.

D. *Notable work in memorability prediction*

Using a convolutional neural network Baveye et al. were able to improve an image memorability score compared to the previous work conducted in the field [3]. Similarly to how the memorability score was calculated in the Medieval memorability task, participants were shown a series of images and then asked if they could recall them after a period of time. These recall rates were then used as the memorability score. They found that emotional negative biases impact image memorability. The more a negative emotion was portrayed on people's faces, the more memorable that image was.

This emphasised the importance of controlling for this with images distributed evenly across 'emotional space'. Thus, the negative emotional bias could be seen as an indicator that memorability is linked to facial features. Further to this in a 2017 paper, Shekar S. constructed a prediction model for video memorability [17]. They demonstrated that a multitude of features play an important role in this prediction, namely; Saliency, a measure of interestingness in videos, using pixel level features but also eye tracking databases with heat maps of interesting sections of pictures. This method had been established in 'Learning to predict where humans look' [9] which emphasised video semantics, i.e. captions associated with the videos and C3D or convolution 3D, which returns generic features associated with a video based on a 3D convolutional network [22].

E. *Notable participants from MediaEval 2018*

Linear Models for Video Memorability Prediction Using Visual and Semantic Features

[8]. The aim of this paper was to extract some of the provided features in the MediEval data set such as Hierarchical Matching Pursuit (HMP) and C3D as they were easily extractable and concatenate colour histogram and local binary patter (LBP) also including the provided caption in the data set.

From their feature extraction, the researchers ran a series of linear models over the data to determine a probability score for memory. This paper achieved the highest short and long term memorability score in last year's competition.

'Predicting Media Memorability Using Deep Features and Recurrent Network' [23] aimed to examine sequential frames within the video data set to predict video memorability. It focused on the temporal factors rather than an ensemble or some of the other features present in the data set. In contrast to the *winning* or best performing work, it conducted some more complex feature engineering and extraction. Rather than using C3D, they split the videos into 8 frames and significantly reduced the computational load. These frames once extracted input into a pre-trained convolutional network. From here the frames become an input of the long short-term memory (LSTM), an artificial recurrent neural network.

II. Neuroscientific justification for using faces to predict memorability

A. *Why We Evolved to Recognise Faces*

Humans are inherently social creatures [6]. In order to function in social groups, it is advantageous to be able to distinguish members from one another. It helps us to differentiate between a threat and a friendly encounter, also helping us to form long-lasting relationships that ultimately lead to self-preservation.

Faces also present important stimulus for differentiating social status such as age, gender and the current emotional state of humans and primates. Understanding this information is vital for navigating social structures and how to act in reciprocal relationships and environments [15]. Furthermore, there is a wealth of evidence to suggest that faces and facial recognition inform and influence mate selection and sex judgements in primates. This influence occurs by facial signals like age, attractiveness or health and trustworthiness. Each are important factors to consider in sexual selection. Paar concluded that the recognition of faces goes beyond mere identity recognition and that humans evolved to use this function in a multitude of ways [15].

Humans also evolved to use face recognition for threat perception. If there are elements of faces that indicate there is a potentially life-threatening risk present in an approaching person, recognition of these elements will lead to self-preservation and greater longevity, both favourable characteristics to pass on to progeny. We recognise this threat through emotional expression in the face but also in recognising non-stereotypical faces. Non-stereotypical faces can be thought of as deviations from our group's normal. [10]

B. *Neurological Mechanisms behind Facial Recognition and Memorability*

The neurological mechanisms of facial recognition predate our cortical systems and have existed for millennia. Cortical systems are responsible for higher order functions and were the most recent to evolve but facial recognition mostly resides in the temporal and occipital lopes just above the cerebellum shown in Figure 1. Having evolved prior to the areas of the brain that are responsible for higher order and intellectual function, these lobes and thus facial recognition have played an important role in how we as humans evolved [21].

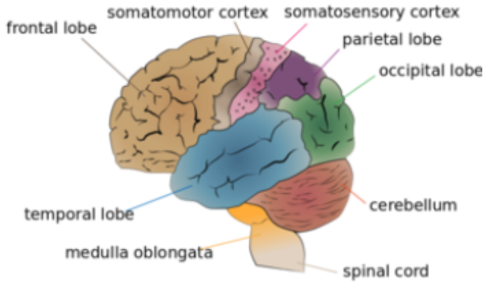


Figure 1. Temporal and Occipital lobes

While no one area of the brain is usually wholly responsible for a function, the temporal and occipital lopes play a key role in pattern detection and facial recognition and commitment to long term memory [18].

The path of memory creation takes places through exposure to a stimuli (like auditory or visual), encoding that information to then create a construct within the brain. This construct is then accessed in short term memory and subsequently 'written' into long term memory after repeated exposure [12].

Face detection differs from other object detection in humans. For example there is an entire cortical region of the brain called the fusiform face area which is dedicated to recognising faces and thus faces are recognised as a whole rather than the sum of its parts like other objects. This means memory creation requires less energy and resources from the brain. Humans are better and more adapted to recognise faces ahead of other objects. This faster recognition also allows for faces to be committed to long term memory with less effort [13].

C. *How face recognition translates to the MediaEval task*

Understanding how and why faces are memorable is of importance to predicting what makes a video memorable. As demonstrated above, there are numerous biological and neuroscientific reasons that humans have adapted to recognise faces quickly but also commit them to memory. Based on this intuition, it is reasonable to evaluate and experiment through machine learning how the presence of a face in a video impacts memorability prediction in the MediaEval task.

The current data set does not explicitly incorporate or exclude faces in the videos so creating features that includes this data will likely have a positive impact on memorability score prediction by discriminating between the presence and absence of faces, according to the theory outlined earlier. It's also worth noting that the participants who were asked to view and subsequently recall if they had seen a given video in the past had no emotional relationship or connection with the people and faces presented in the clips. The participants have an emotional neutrality to the subject and therefore this will not bias the memorability scores.

There has also been some previous work completed in computer vision and experimental psychology that motivates us to explore how the presence of faces in videos could improve memorability prediction. Isola outlined the presence of a face in a photo contributes positively to its memorability [13]. Similarly, it was discovered that not only did it improve memorability after a single exposure but also that this improvement is consistent across all the observers when asked if they could recall an image after exposure [4]. Based on this, it seems reasonable to test the hypothesis that the presence of a face will increase a video's memorability. That is what we set out to provde in this work.

III. Related Technical Work

This section will explore the existing features in the MediaEval data set in more detail along with describing some machine learning architectures that will be used. Understanding their calculation, relevance and prior contribution to image and video memorability prediction is useful to help prioritise feature selection and development of new features. These include image-based features extracted from specific frames, video-based features that include the temporal and sequential aspects of the video, captions-based features and transfer learning features inherited from previous work in computer vision.

A. *Frame based features*

1) Histogram of Oriented Gradients (HOG): HoG descriptors are useful for both edge and object detection. They are computed by calculating a single feature vector for the entire grayscale frame. It's useful to think of the creation of this vector as sliding a window across a frame (i.e. a pixel grid) and the descriptor being calculated for each position in the image [7]. The gradients in each of these positions can be thought of as the change intensity in that area of the frame.

This generates a gradient vector which is then split into angular bins and represented in a histogram. Each of these bins can reflect the gradient magnitude. This reduces the size of the feature vector that is stored but maintains the useful information within. The gradient of each angular bin will show the magnitude of intensity change in each direction. For example in figure 2, the white dashes represent this intensity change. To then detect if a particular object is present in the image or not, support vector machines (SVM) are used as a popular choice of classifier.

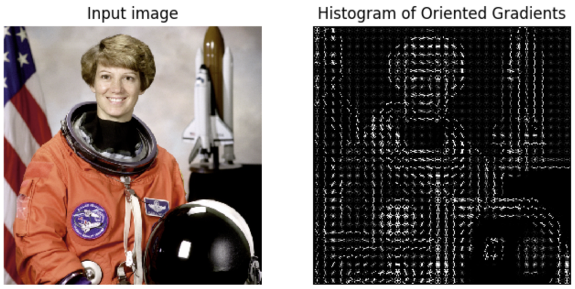


Figure 2. HoG Descriptor

2) Local Binary Pattern (LBP): LBP is another feature that can be used in edge and object detection but is also useful for evaluating textural changes within frames [1].

The LBP takes a window of pixels and converts it into a single value by comparing every neighbouring pixel with the central pixel in a 3x3 grid [1]. This comparison yields a luminosity value; if the pixel's luminosity value is greater than or equal to the centre pixel's value, mark the pixel as a 1 otherwise set it to 0. This grid is then converted into a number, a series of 8 bits to be precise. This can be converted into the decimal number system and used to train a machine learning algorithm to recognise objects, patterns and gradient changes. This pattern is invariant to illumination because the distances between the pixels' illumination will remain constant if the lightness changes. Transitions from areas with a lot of ones to a lot of zeros will show parts on of the entire frame going from light to dark and help detect edges.

In summary, a histogram of LBP values will show intensity, changes in regions will show edges and the machine learning (ML) will piece these together to show objects and adding a temporal dimension will show changes from frame to frame providing useful information about how a particular scene develops.

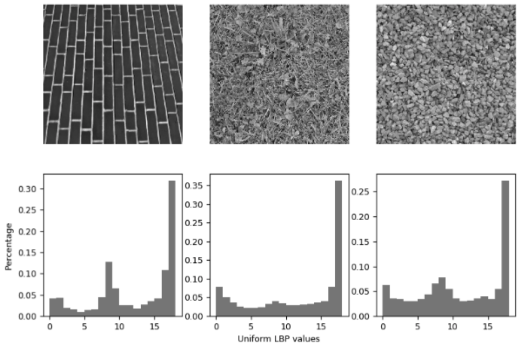


Figure 3. Local Binary Pattern (LBP) values for 3 images

3) Red Green Blue (RGB) histograms: RGB histograms represent the distribution of colour in an image. They are a useful descriptor that can serve as a feature vector and are particularly useful for computer vision systems that may want to compute the similarity of two images. Figure 4 is an example of colour distribution which has been grouped by pixel count bins

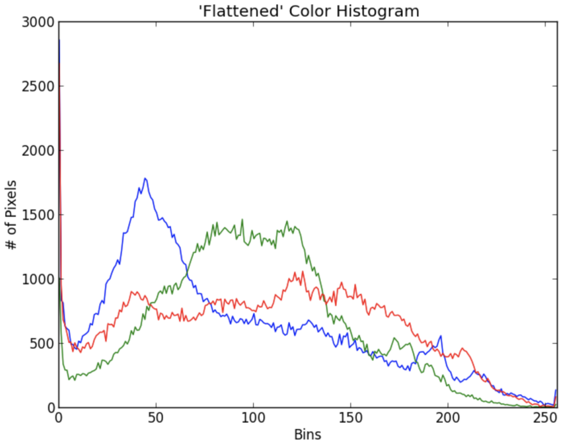


Figure 4. RGB

B. *Video based features*

1) Convolution 3D (C3D): C3D is obtained by training a deep 3D convolutional network on a large annotated video data set. The original data set contained objects, actions, scenes and other frequently occurring categories in videos. Unlike 2D convolution, this network architecture is able to consider temporal information The aim of C3D is to learn spatiotemporal features within the video. "These 3DConvNets encapsulate information related to objects, scenes and actions in a video, making them useful for various tasks without requiring to fine tune the model for each task" [23].

2) Scale-Invariant Feature Transform (SIFT) : SIFT takes particular segments of images and feature matches between a sequence or series of images, in this case the series of images are the multiple frames within a video but the technique is also used in object detection. Frame content is transformed into local feature coordinates that are invariant to transformation, rotation and scale which allows for segment detection between frames [11].

C. *Transfer Learning features*

1) Fully Connected (Fc7) layer from AlexNet: Alexnet is a convolutional neural network that gained notoriety by winning ImageNet Large Scale Visual Recognition challenges in 2012 and contributed to the recent popularisation of deep learning [2]. Figure 5 details an input image on the left-hand side passing through a series of convolutional layers in order to learn patterns from the input image. The final dense or fully connected layers of the network make the classification decision. The transfer learning takes place when these final fully connected layers are retrained on the memorability scores from the MediaEval dataset in order to improve memorability prediction.

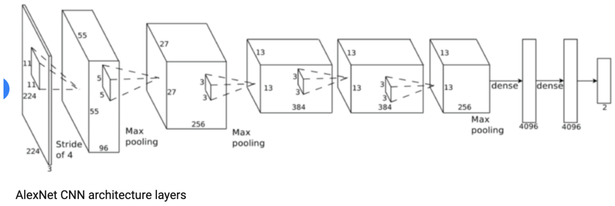


Figure 5. Fc7 from AlexNet

2) Haar Cascades for Face Detection: Haar cascades are a machine learning approach that are trained on many positive and negative examples of a particular object. When it is shown a new instance, the Haar classifier will be able to detect if the object is in the image. In more detail Haar cascades are used for the classification of objects like faces by concatenation of features from a number of weaker classifiers. Haar features are computed by calculating the difference between white and black pixels in a grayscale image, combinations of these patterns are then used to build Haar features. The pixel size of images is then varied in size in order to make the object detection size invariant [24].

Machine learning models are then trained on the multitude of positive and negative examples associated with a particular object. New images can then be passed through this cascade to determine if a relevant object is present or not. Figure 6 shows how the Haar features are extracted from an image, then used to train a model of positive and negative examples, finally being used to detect faces in images on the right.

D. *Captions features*

The MediaEval data set comes with pregenerated captions to describe the videos. This is an example of one video's captions: "blonde-woman-is-massaged-tilt-down". As the data

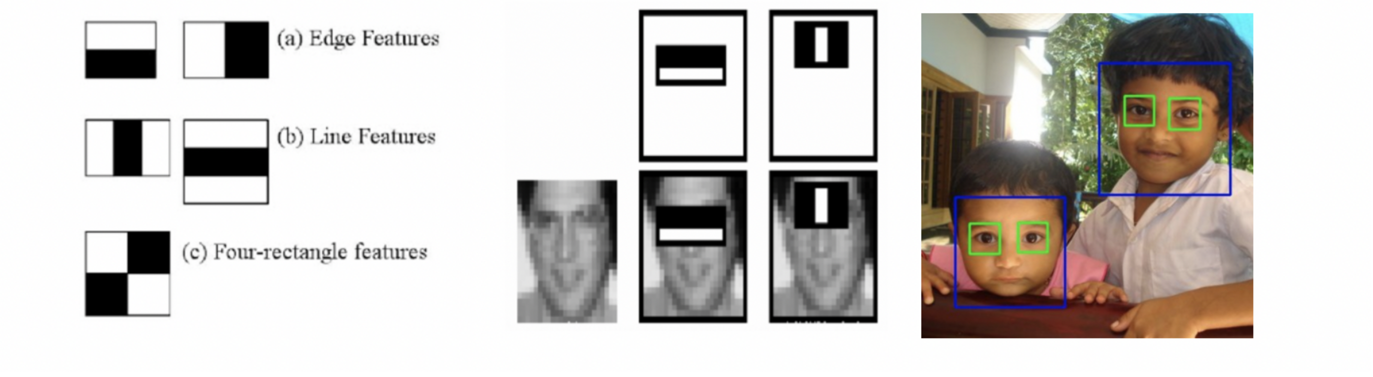
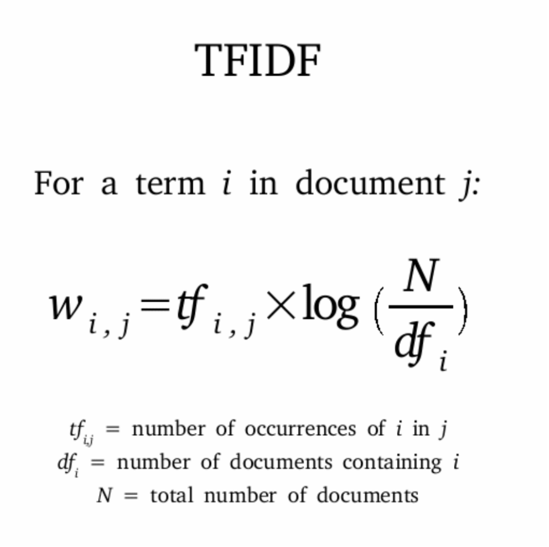


Figure 6. Haar Cascades

is not in numeric form, some natural language processing (NLP) techniques can be applied to convert these captions to something that a model can work with.

1) Term frequency, inverse document frequency (TF-IDF): TF-IDF vectorization is a method that converts each word into a numerical value and applies a heavier weight to more important and less frequently used words, so words appear less frequently will have a higher score and vice versa. It is calculated by taking each word in a set (term frequency), multiplied by a measure of how significant a particular word is or how rare (as words that appear less often contain more information) that word is relative to each caption in the set.

2) One Hot Encoding : One hot encoding is a processing method that can be applied to categorical features, and can be deployed in NLP. To do this, we create a matrix with each distinct word from the captions represented as a column vector and each video represented as a row vector. Whenever a word is present in a video's caption the corresponding position in the array will be 1, otherwise 0. The ensuing matrix will contain mostly zeros but can now be used as a model's feature.


TFIDF

For a term i in document j:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Figure 7. TF-IDF

E. Model Architectures For Experimentation

1) Convolutional Neural Networks (CNN): CNNs are a type of artificial neural network, which consists of multiple layers and interconnected nodes. The initial nodes of the input layer are connected by initial weights and inner nodes connected by an activation function. The output nodes contain the prediction of the class of an input or an associated probability.

These networks are trained using back propagation, a process by which the error is calculated between a predicted value and ground truth. This error is passed back through the network in order to optimise or reach a minimum to the error function. Gradient descent is a mathematical method used to adjust the weights and biases in an optimal manner and uses differential equations to optimise the system [16]. Both three and five hidden convolutional layers were used with combinations of sigmoid and relu activation functions. Using five hidden layers had no material difference in performance results so after initial experimentation the three-layer architecture was used. Dropout layers were also deployed in an attempt to minimise overfitting on the data set. This work utilizes relu (all or none 'firing') and sigmoid (continuous 'firing') activation function.

2) Support Vector Regression (SVR): This is a variant of support vector machines (SVM) used for predicting continuous values rather than classification tasks. The core idea between the two methods are the same; to minimise error, individualising the hyperplane which maximises the margin. SVR calculates linear regression in higher dimensional space and is therefore useful in machine learning but rather than try to minimise the error like in linear regression, SVR fits the error within a certain threshold which is the boundary line [19].

3) Bayesian Ridge Regression : Bayesian ridge regression model formulates a linear regression by using probability distributions rather than point estimates. The aim of Bayesian ridge regression is not to find the single "best" value of the model parameters, but rather to determine the posterior distribution for the model parameters. Both the response and model parameters come from the probability. Posterior probability is dependent on conditional values associated with the data [16].

IV. Data Preparation and feature engineering

This section outlines some of the technologies and programming libraries used to extract, explore and engineer new features from the data set. An emphasis will be placed on how theory-based hypotheses can be transformed into experiments using analytics.

A. *Overview of the Data Set*

As mentioned previously the data set was comprised of a number of pre-computed features from 10,000 (soundless) short videos extracted from raw footage split into 80/20 train and test sets. Each video is associated with two scores of memorability that refer to its probability to be remembered after two different durations of memory retention [5]. These features were collected from MediaEval using a file transfer protocol (FTP) connection and stored in Google Drive. The data contained pre-computed features, meta data and the raw source videos.

The processed features like C3D and Fc7 layer were stored in text and CSV files as comma separated features or with key value pairs. In order to model these features they had to be transformed into a pandas data frame to be manipulated and explored.

Figure 8 shows the distribution of ground truth memorability scores. Most noticeably short term memorability scores have a higher mean and lower variance. Which is interestingly analogous to how humans process memory; we have more short term memories over our lifetimes than long term.

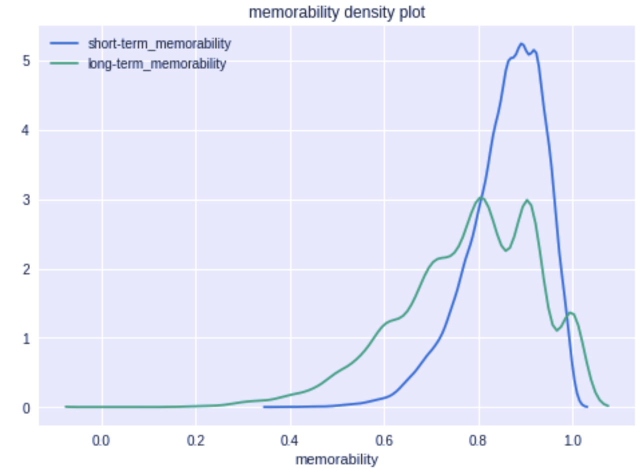


Figure 8. Memorability Density Plot

Figure 9 details the map of correlations between the short term ground truth scores and the long term. It clusters in the top right, which again would make sense that, if a video had a higher short term memorability score, it would also have a higher long term score.

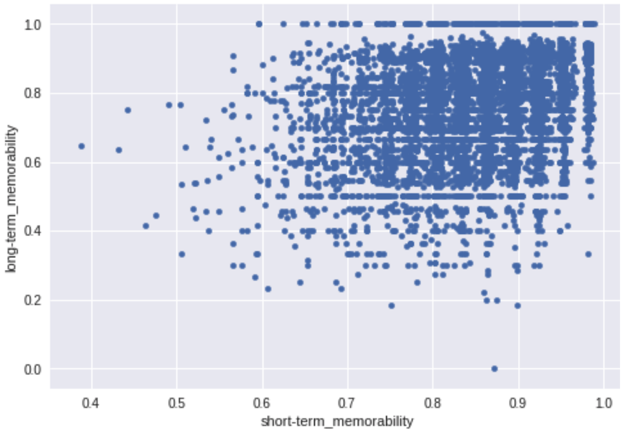


Figure 9. Memorability Correlation Scatter Plot

B. *Computing environment and challenges*

The entire practicum operated in Google Colab, a Jupyter notebook environment which provisions resources from Google's data centres and provides free run time in the Cloud. Using Colab made sense because it allowed the work to place more emphasis on building a machine learning pipeline to take raw unprocessed data to create features and eventually generate memorability predictions rather than provisioning and maintaining virtual machines to execute code.

The environment offers free central processing unit (CPU), graphics processing unit (GPU) and tensor processing unit (TPU) runtime. During initial analysis the CPU performed significantly slower than GPU and TPU for both training neural networks and other machine learning tasks. For file processing while extracting and engineering new features, the CPU also frequently hits its limits. Comparative differences between GPU and TPU were not significant during this research but for larger processing and bigger operations like computing face features the environment interrupted the runtime to prevent misuse of the service.

C. *Engineering Face Features*

From a processing perspective, videos contain a lot of information and present some challenges when computation resources are limited. Common video processing and computer vision libraries like OpenCV also treat videos like a series of frames or images and conduct any processing on each frame in a video. This method can be problematic because a 7-second video can have more than 200 frames which can quickly become computationally expensive. The solution implemented to circumvent this issue was to use key frame extraction. This is a process whereby a video is reduced by extracting a subset of the total frames to reduce both the time and resources required to engineer a feature using face detection. This was achieved by extracting a frame per each second of the videos.

Once the key frames were extracted and stored in a folder, a number of functions were written to turn the frames into a feature that could be stored in a analytics base table containing a number of features. Using Open-CV, a C++ computer vision library, the key frames are converted to grayscale and stored in an array. The next stage is to use a pre-trained Haar Cascades file which contains facial features encoded into XML files. The grayscale images are then compared to the XML and if it meets a threshold of face criteria the face detector will return a positive result. The criteria and thresholds were altered to improve accuracy before the feature generation was finalised.

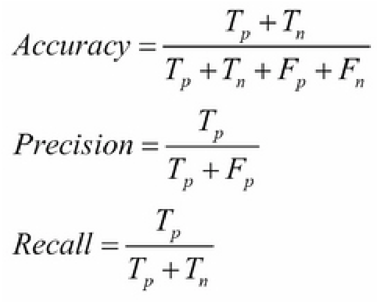
The next step in the process is to read the output in the face detector for each frame and write the result to a pandas' data frame. Some data manipulation is subsequently used to calculate the feature value for each series of frames in a video. The first calculation method was to take the number of frames in which a face appeared and divide it by the number of total frames in a video. The second calculation methodology used was a simple Boolean value: if a face was present in a series of frames then the feature value is set to 1, otherwise 0. The second calculation methodology yielded stronger performance when used to predict memorability scores, so from here on this will be referred to as the facial feature.

D. *Evaluating Face Detection Accuracy*

Like any detection system, the process for detecting a face in a video and creating a feature will not perform perfectly under every circumstance. To evaluate the performance of the system 10% of the training set was sampled (200 videos in total) and manually tagged if a face had been accurately detected in the frames or not. From this manual tagging the below standard evaluation metrics were calculated.

E. *Evaluation Metrics*

- Accuracy details the ratio of correct prediction over the total number of observations
- Precision is a measure for correctly predicted positive observations of total number of predicted positive.
- Recall is a sensitivity measure, correctly predicted true over everything that is true.



$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

$$Precision = \frac{T_p}{T_p + F_p}$$

$$Recall = \frac{T_p}{T_p + T_n}$$

Figure 10. Definitions of Accuracy, Precision and Recall

V. Experimentation and Analysis

Results are obtained from the test data and compared to the ground truth memorability scores. They are also split into two distinct phases. The first phase used the existing set of features to test the machine learning pipeline on the data set but also to set up prediction benchmarks to evaluate the improvement of face detection. The second stage then used the newly engineered facial features to evaluate any change or enhancement in memorability prediction.

A. *Convolutional Neural Net with Captions*

The rationale for starting with captions using one hot encoding was based on the MediaEval 2018 winning paper which deployed a linear model using captions [5]. Using a convolutional neural net as described above yielded positive results achieving a Spearman Rank Correlation of 0.273 to the ground truth for short term memorability prediction and 0.151 for long term memorability. These correlations are calculated between the ground truth memorability score and the model's predicted memorability score associated with each video.

B. *Ensemble of Captions and Fc7 layer*

Using an ensemble approach with the captions' CNN and the pre-computed fully connected layer from AlexNet resulted in a poor performance, scoring a short-term memorability correlation of 0.035 and long term memorability correlation of 0.009. This architecture was the most conceptually complex as it used the prediction layer from the captions network as an input layer, along with the fully connected layer from AlexNet to then obtain predictions. While this approach was interesting, it performed the worst out of all the models and techniques used.

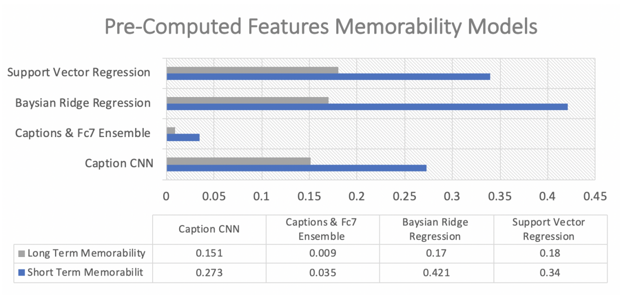
C. *Bayesian ridge regression using C3D, SIFT and Captions*

Using a combination of the video based and pre-computed features including one hot encoding yielded the best performing results of any feature combination.

Bayesian regression aims to formulate a posterior distribution for a given set of inputs [18]. In this instance the prior probability was calculated under the frequentist definition using the ground truth score (averages across participants for memorability). Bayesian regression then formulates posterior probability distributions for the inputs i.e. C3D, SIFT and Captions. This resulted in a short term memorability correlation of 0.42 and long term of 0.17.

D. *Support Vector Regression (SVR) using C3D, SIFT and Captions*

SVR resulted in some strong predictive scores with 0.34 correlation to short term ground truth memorability and 0.18 to long term. Interestingly, using the same features as the



	Caption CNN	Captions & Fc7 Ensemble	Baysian Ridge Regression	Support Vector Regression
▨ Long Term Memorability	0.151	0.009	0.17	0.18
■ Short Term Memorabilit	0.273	0.035	0.421	0.34

Figure 11. Existing Feature Model Spearman Scores

Bayesian ridge regression and only differing in the model architecture led to different results.

E. Incremental Facial Features

The addition of the facial features was then used and benchmarked against, which provided some interesting results. The first result used Bayesian regression which was the best performing architecture using pre-computed features. Using face detection in a single feature model produced poor results with a short term memorability correlation of 0.05 and long-term memorability of 0.01. The second approach for the face detection involved the use of both the CNN with captions and face features. This approach was more successful and resulted in increased memorability prediction for both the short and long term scores with 0.36 and 0.27 respectively. This represents an increase of 13.75% in short term memorability and 22.76% in long term memorability compared to the CNN with captions only. The next section will explore the differences between modelling with and without facial features.

F. CNN w/ Captions Vs. CNN w/Captions + Faces

Analysing the results of both model architectures provided some interesting insight into media memorability. On average the results improved but this was not the case for every video. In an effort to explore the differences between these videos it is useful to understand how the individual video memorability predictions changes between the two models. For this, it helps to analyse the distributions and relationships of each model's error. The error in this instance is the difference between the ground truth memorability score in the test set and the corresponding model prediction for each video.

When it comes to the error a reduction indicates that our prediction for a specific video is better. Each red dot in Figure 12 represents the error of a video using the CNN that incorporated facial features, while the blue marks represent the CNN prediction error without faces. The lines in the graph represent an ordinary least squares regression line of best fit. There is a distinctly noticeable decrease in error between the blue and red clusters, so adding faces as a feature has some clear marginal benefit.

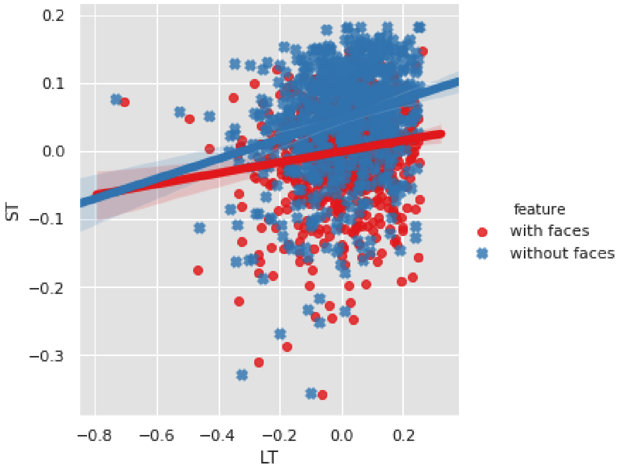


Figure 12. Linear Regression of Error Terms

The errors from both these models are further explored in figure 13 by plotting the distribution of short and long term memorability error. Again, the ST shows a sizeable drop in error compared to before facial features are added. This difference is less evident with long term memorability error.

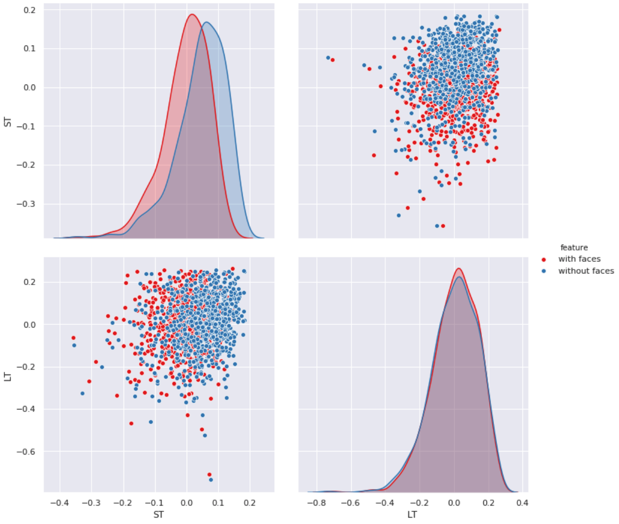


Figure 13. Distribution of Errors

VI. Understanding Where Face Features Helped

Through exploration of individual observations where memorability prediction improved some fascinating insights were uncovered. The largest increases existed in two distinct types of videos; firstly, a large increase in memorability could be identified following changes in emotion between the frames, for example if frame no.1 showed a happy face and the final frames expressed sad or angry faces. Secondly, an increase in memorability also existed in instances where faces appeared to be threatening or dangerous within a specific video.

Interestingly, these findings are both aligned with the existing neuroscientific literature. As humans are more likely to detect and thus remember angry faces [14], faces displaying anger or changes from positive to negative emotions are detected more efficiently [3]. These findings are also confirmed in fMRI, where experiments have showed higher response for those cases than neutral and unchanging facial expressions [20].

Conversely there was also some consistency in the videos where prediction decreased or remained level, if the face detection triggered a false positive and the feature incorrectly expressed the presence of a face. Furthermore, if a face was not present in the video, the addition of the facial feature resulted in little to no change in memorability prediction as expected.

VII. Conclusions & Further Study

A. *Conclusion*

Video is taking up a larger proportion of both internet traffic and time people spend online, understanding what makes a video more or less memorable is of high importance.

This work sought to explore the features which lead to videos being more memorable and created new features to enhance memorability. Through evidence-based inquiry and experimentation, the presence of faces was shown to increase video memorability.

The MediaEval Media memorability task involves trying to predict a ground truth memorability score associated with 10,000 seven second videos. The contest organisers included a set of pre-computed features to help in this prediction task. Through exploration of these pre-computed features, this work was able to achieve a strong memorability score using Bayesian ridge regression, including a Spearman rank correlation to the ground truth short term memorability score of 0.42.

However, the foundation of this work involved exploring the neuroscientific rationale behind facial recognition in humans and how it could improve memorability prediction, engineering and preprocessing new facial features and finally experimenting with these engineered features using a benchmark model for comparison.

Humans have evolved to detect and remember faces more vividly, ultimately to increase self-preservation. Humans do this to behave and function effectively in social environments. Further to this, recognising and remembering threatening facial expressions is conducive to survival [10].

Based on this evidence, engineering facial features to predict memorability was a valid avenue to explore. Once these features were explored, they were modelled and compared to a benchmark set of features using the same convolutional neural network architecture. The results showed 14% increase in short term memorability prediction and 23% improvement in long term memorability prediction.

Faces provided a significant lift in memorability prediction when compared to a benchmark and on average the presence of faces in video made them both more memorable and enhanced memorability prediction. This effect was most extreme where there was threatening facial expressions in videos or where the emotions in videos changed from positive to negative i.e. from happy to sad or angry.

B. *Further Study*

The videos that showed the largest error reduction in memorability prediction all shared some common characteristics. There was either some emotional complexity in the image, i.e. the emotions expressed on people's faces changed throughout the clip or there was some emotional escalation or something that could be perceived as a threat. This was an interesting finding as it's well documented in neuroscience that angry and threatening faces are remembered faster and more vividly [14] and this persists across various ages. The reasons for this are in line with the self-preservation argument mentioned above [10].

Computationally exploring emotional escalation and threat perception on memorability more in-depth would be useful to improve the understanding of video memorability. Improving this understanding would also further enhance memorability prediction.

References

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. Number 12, pages 2037–2041. IEEE, 2006.

[2] SH Basha, Shiv Ram Dubey, Viswanath Pulabaigari, and Snehasis Mukherjee. Impact of fully connected layers on performance of convolutional neural networks for image classification. 2019.

[3] Yoann Baveye, Romain Cohendet, Matthieu Perreira Da Silva, and Patrick Le Callet. Deep learning for image memorability prediction: The emotional bias. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 491–495. 2016.

[4] Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. Intrinsic and extrinsic effects on image memorability. volume 116, pages 165–178. Elsevier, 2015.

[5] Romain Cohendet, Claire-Hélène Demarty, Ngoc Duong, Mats Sjöberg, Bogdan Ionescu, and Thanh-Toan Do. Mediaeval 2019: Predicting media memorability task. 2019.

[6] Kim M Curby and Isabel Gauthier. A visual short-term memory advantage for faces. volume 14, pages 620–628. Springer, 2007.

[7] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. 2005.

[8] Rohit Gupta and Kush Motwani. Linear models for video memorability prediction using visual and semantic features. In *MediaEval*. 2018.

[9] Daniel Kang, John Emmons, Firas Abuzaid, Peter Bailis, and Matei Zaharia. Noscope: optimizing neural network queries over video at scale. volume 10, pages 1586–1597. VLDB Endowment, 2017.

[10] Heather M Kleider-Offutt, Alesha D Bond, Sarah E Williams, and Corey J Bohil. When a face type is perceived as threatening: Using general recognition theory to understand biased categorization of afrocentric faces. volume 46, pages 716–728. Springer, 2018.

[11] David G Lowe et al. Object recognition from local scale-invariant features. In *iccv*, volume 99, pages 1150–1157. 1999.

[12] Matthew MacDonald. Your brain: The missing manual: The missing manual. " O'Reilly Media, Inc.", 2008.

[13] Alex Martin, James V Haxby, Francois M Lalonde, Cheri L Wiggs, and Leslie G Ungerleider. Discrete cortical regions associated with knowledge of color and knowledge of action. volume 270, pages 102–105. American Association for the Advancement of Science, 1995.

[14] Mara Mather and Marisa R Knight. Angry faces get noticed quickly: Threat detection is not impaired among older adults. volume 61, pages P54–P57. Oxford University Press, 2006.

[15] Lisa A Parr. The evolution of face processing in primates. volume 366, pages 1764–1777. The Royal Society, 2011.

[16] Shashi Sathyanarayana. A gentle introduction to backpropagation. 07 2014.

[17] Sumit Shekhar, Srinivasa Madhava Phaneendra Angara, Manav Kedia, Dhruv Singal, and Akhil Sathyaprakash Shetty. Techniques for enhancing content memorability of user generated video content. Google Patents, October 31 2017. US Patent 9,805,269.

[18] EE Smith and SM Kosslyn. Cognitive psychology: Mind and brain. 2007, 2007.

[19] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. volume 14, pages 199–222. Springer, 2004.

[20] MM Strauss, Nikos Makris, I Aharon, Mark G Vangel, J Goodman, David N Kennedy, GP Gasic, and Hans C Breiter. fmri of sensitization to angry faces. volume 26, pages 389–413. Elsevier, 2005.

[21] Sathesan Thavabalasingam, Edward B O'Neil, Jonathan Tay, Adrian Nestor, and Andy CH Lee. Evidence for the incorporation of temporal duration information in human hippocampal long-term memory sequence representations. volume 116, pages 6407–6414. National Acad Sciences, 2019.

[22] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497. 2015.

[23] Duy-Tue Tran-Van, Le-Vu Tran, and Minh-Triet Tran. Predicting media memorability using deep features and recurrent network. In *MediaEval*. 2018.

[24] Phillip Ian Wilson and John Fernandez. Facial feature detection using haar classifiers. volume 21, pages 127–133. Consortium for Computing Sciences in Colleges, 2006.