# Video Memorability Prediction Based on Visual Features and Captions: MediaEval 2018

**Apurva Saswadkar**
*MSc in Computing*
*Dublin City University*
*Dublin, Ireland*
apurva.saswadkar2@mail.dcu.ie

## ABSTRACT
Memorability is the ability to remember anything that is worth remembering. A huge chunk of image and video media data is generated every day on social and other platforms. This paper focuses video memorability prediction which is a part of challenge given by MultiEval [1]. In this paper, we explore different machine learning models to predict how memorable was a video to a viewer. Our approach includes the implementation and comparison of different models to find out the most reliable model that gives the highest accuracy for short term and long-term memorability.

## 1. INTRODUCTION

Memories can be categorized as 'Short-term' and 'Long-term' because some memories tend to evolve with time and some memories fade. A significant part of human cognizance is the capacity to recollect and review photographs and video recordings. Video memorability is also classified as a short-term and long term that can be predicted based on the memorability score which states the probability of remembering measure of detail. In the given work we investigate the different video, image features of the video along with captions to predict the memorability score. In the given dataset, many image and video features are provided such as HMP (Hierarchical Matching Pursuit), C3D, InceptionV3, Colour Histogram etc. we have conducted extensive analysis on three features among all the given features. To train the models we have used C3D and captions because according to the results of previous work done, captions individually have given more accurate results compared to the other features [4,6].

Using two features we train different models to predict memorability score and these models are further evaluated based on the Spearman's correlation coefficient. we have implemented SVR (Support Vector Regressor), RF (Random Forest), Decision Tree Regressor, LASSO regression analysis models on captions and C3D feature. All these models give Spearman's correlation coefficient for both 'short-term' and 'long-term' memorability based on which the best model is determined.

## 2. RELATED WORK

The concept of human memorability has been studied extensively before by psychologists and neurologists where their main focus was visual memory which means a person's ability to remember details of images and videos [2]. In Computing, initially, the computational studies were conducted on cognitive matric and memorability of Images where it was concluded that the memorability of an image is consistent irrespective of subject and context. This clearly stated that image memorability is an intrinsic property of images [3]. The research and work in case of video memorability has been recently begun to interest and involve more people. The great initiative was taken by MediaEval for multimedia evaluation in 2018 where the participants were challenged to use high-level visual features, and image/video captions for memorability prediction. The major finding from the research was that the models trained using captions only gave the best results [4,6]. In previous work, different models were implemented on aesthetic and semantic features and it was found that the semantics were more effective and relevant to both short term and long-term memorability [5].

## 3. DATASET DESCRIPTION

Dataset plays a really important role in any analysis. In this case, as the task of video memorability prediction is performed as a part of a MediaEval challenge, we have worked on the dataset of 8,000 short soundless videos. These videos are given by the MediaEval:2018 with the license under which use, redistribution and processing of those videos is permitted [1]. For training and testing, this data set of 8000 videos was further split into 6000 videos as development dataset and 2000 videos as test dataset. All of these videos were taken from the footage created by professionals while making content [1]. The Development data contains sources, pre-extracted visual features of videos like C3D, HMP, colour histogram along with captions of those short videos. It also contains Ground truth CSV files which include the name of the video, short-term score, long-term score and annotations used to calculate short-term and long-term memorability score. The test data set consists of all the features and files as development data set except for the ground truth files.

## 4. APPROACH

The approach we have followed for memorability score prediction is training various models on one visual feature along with a caption and then testing those models to get the memorability score. The Features and Model used are as follows:

### 4.1 Visual Features

Following features were used as the input for training different machine learning models:

- C3D Feature: C3D is a feature that gives the final classification layer of the C3D model. It is a video specialized feature with a dimension of 101.

- Captions: Captions was chosen as it was found as the feature giving better memorability score as compared to other features in previous studies. Captions used were the textual description of videos present in the dataset.

The reason for choosing the C3D feature is that it was the better performing feature of video specialized feature that is it gives better accuracy as compared to HMP. Captions are selected as the mandatory feature as it gives the best memorability score among other features as mentioned in the previous papers.

### 4.2 Models Implemented

To predict the video memorability, as mentioned above we implemented SVR, RF and LASSO deep learning models to derive the memorability score.

- Support Vector Regression (SVR)

SVR gives us the flexibility to define how much error is acceptable in our model [4]. Our data-set contains videos that have different annotations and we need to find the effect of number of annotations on prediction. Keeping this nature of SVM and our purpose of memorability prediction in mind, we have trained the SVR model on the captions and C3D visual features of the videos in development dataset to predict the memorability score of the data in testing dataset.

- Least Absolute Selection and Shrinkage Operator (LASSO)

It is a regression method that improves the prediction accuracy by performing variable selection and regularization simultaneously. This model uses a penalty that affects the value of the regression coefficient. Regression coefficient moves inversely with the penalty. It implements the L1 Normalization method that uses the tuning parameter as the amount of shrinkage [8]. According to the previous work done, the LASSO model gives the best accuracy for HMP and colour histogram feature.[7] In my approach, we have trained and tested the LASSO model on captions and C3D features of training and testing data.

- Random Forest (RF)

Random forest is a supervised machine learning algorithm that learns from labelled data and gives predictions based on learnings. RF uses decision trees as the base, it creates a bunch of decision trees and makes predictions by combining the outcomes of decision trees. The Combining of the model and outcome is done through an ensemble method using a voting mechanism. So, the outcome of each parallelly executing decision tree is collected and the most common outcome of all the outcomes is selected as the final outcome. As we can see this method uses decision trees which are generally used to classifying both numeric and non-numeric data type it is ideal for our research. The reason for this is the features that we have selected i.e. captions and C3D where one data is numeric and the other is non-numeric. Hence this model will give better result on our data and help in classifying the memorability of the videos.

## 5. Results and Findings

The below table contains the results obtained by implementing SVR, LASSO and RF on captions, C3D features for predicting video memorability.

| Features / Models | Caption | C3D | Caption + C3D |
|---|---|---|---|
| **SVR** | 0.421 | 0.038 | 0.402 |
| **RF** | 0.423 | 0.015 | 0.334 |
| **LASSO** | 0.417 | 0.015 | 0.399 |
| **SVR + LASSO** | 0.454 | 0.019 | 0.329 |
| **SVR + RF** | 0.454 | 0.015 | 0.403 |
| **LASSO + RF** | 0.449 | 0.008 | 0.412 |
| **SVR + LASSO + RF** | 0.465 | 0.009 | **0.431** |

Table 1. Models and their Short-Term memorability prediction accuracy obtained for C3D and captions

| Features / Models | Caption | C3D | Caption + C3D |
|---|---|---|---|
| **SVR** | 0.187 | 0.019 | 0.198 |
| **RF** | 0.186 | 0.033 | 0.062 |
| **LASSO** | 0.163 | -0.004 | 0.156 |
| **SVR + LASSO** | 0.195 | -0.000 | 0.188 |
| **SVR + RF** | 0.197 | 0.034 | 0.178 |
| **LASSO + RF** | 0.191 | 0.031 | 0.151 |
| **SVR + LASSO + RF** | 0.201 | 0.013 | 0.182 |

Table 2. Models and their Long-Term memorability prediction accuracy obtained for C3D and captions

In the above tables, the first table gives the Spearman's Correlation Coefficient obtained for Short-term Memorability and the second table gives the same for Long-term Memorability. Each row gives the correlation values obtained by implementing models on captions individually, C3D individually and captions, C3D together. These values give the correlation between actual values and predicted values. So the accuracy of the model can be determined by the correlation, higher the correlation, more the accuracy of the model. We have implemented SVR, LASSO and RF models individually and then we have also obtained combined predictions of two and three models to get the better results. It can be clearly seen that all models give higher accuracy for short term memorability compared to the long term memorability.

## 6. Conclusion

The result of this research suggests that different model perform in different ways for different features. It can be seen that the Caption increases the accuracy of the models compared to C3D. In fact, it gives better results than C3D and Captions combined for all the machine learning models. In the case of Short term memorability, all three model combine gives a better result than each model individually. Although this scenario changes in case of Long Term memorability where SVR gives a better result than all models combined.

## REFERENCES


[1] Romain Cohendet, Claire-Hélène Demarty, Ngoc Q. K. Duong, Mats Sjöberg, Bogdan Ionescu, Thanh-Toan Do.2018. MediaEval 2018: Predicting Media Memorability.

[2] Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. 2008. Visual long-term memory has a massive storage capacity for object details. Proceedings of the National Academy of Sciences 105, 38 (2008), 14325–14329.

[3] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. What makes an image memorable? (2011).

[4] Romain Cohendet, Karthik Yadati, Ngoc Q.K. Duong, and ClaireHélène Demarty. 2018. Annotating, Understanding, and Predicting Long-term Video Memorability. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. ACM, 178–186.

[5] Wensheng Sun, Xu Zhang Saginaw.2018. Video Memorability Prediction with Recurrent Neural Networks and Video Titles at the 2018 MediaEval Predicting Media Memorability Task.MediaEval'18, 29-31 October 2018, Sophia Antipolis, France.

[6] Sumit Shekhar, Dhruv Singal, and Harvineet Singh. 2017. Show and Recall: Learning What Makes Videos Memorable. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2730– 2739.

[7] Rohit Gupta, Kush Motwani.2018. Linear Models for Video Memorability Prediction Using Visual and Semantic Features.MediaEval'18, 29-31 October 2018, Sophia Antipolis, France.

[8] Guo-Xun Yuan, Chia-Hua Ho, and Chih-Jen Lin. 2012. An Improved GLMNET for L1-regularized Logistic Regression. Journal of Machine Learning Research 13, Jun (2012), 1999–2030.