IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 21, Issue 6, Ser. II (Nov - Dec 2019), PP 12-17
www.iosrjournals.org

Video Aesthetic Assessment Using Deep Learning

[1]Mr. Shreyas Joshi, Ms.Athira Menon,[2] Prof. M .V. Phatak
[1]*Student, Computer Department, Maharashtra Institute of Technology, Pune, India,*
[2]*Professor and research guide, Computer Department, Maharashtra Institute of Technology , Pune, India*

Abstract: *Initial studies have shown that automatic inference of high-level video quality or aesthetics is very difficult. In this era of internet that facilitates rapid transmission of images and videos, there is a notable influence on the overall population. This makes the ability to assess the aesthetic value of a video very beneficial to various applications. The image processing and analysis community has, for long, attempted to quantify and rectify video quality at a lower level. At a higher level, the perception often affects our emotion and mood, but there has been little headway made in automatic inferencing of the quality in images that affect mood or emotion. Also, low-level image properties are insufficient to characterize high-level perception of aesthetics. Furthermore, there is a lack of precise definitions, assessment metrics, and test data for this problem, despite being desirable for many applications. In this project, we attempt to clear the cloud on the problem of video aesthetics inference from visual content, by defining problems of interest, target audiences and how they affect the problem at hand, assessment metrics, and introduce real-world datasets for testing. While very limited work has been published so far, we hope that this exposition to the subtleties will encourage more contributions.*

Keywords*: 3D-CNN, Deep Neural Network, Optical Flow, Video Aesthetics, Motion Features, Motion Metrics*

Date of Submission: 18-12-2019 Date of Acceptance: 01-01-2020

I. Introduction:

In today's digital world, we face the challenge of developing efficient multimedia data management tools that enable users to organize and search multimedia content from growing repositories of digital media. Increasing storage capabilities at low prices enable the generation and archival of unprecedented amounts of personal multimedia content including digital images and videos. In case of social media platforms, filtering and re-ranking the videos with a measure of its aesthetic value would probably improve the user experience and satisfaction with the search results. In addition to improving search results, another challenge faced by video sharing sites is being able to attract advertisement to the user generated content, particularly given that some of it is deemed to be "unwatchable", and advertisers are typically reluctant to place their clients' brands next to any material that may damage their clients' reputations. We believe that the aesthetic analysis of such videos may be one of the tools used to automatically identify the material that is "advertisement worthy" vs. not. Here, we focus on building computational models of the aesthetic appeal of consumer videos.

II. Literature survey

Title: Video Aesthetic Quality Assessment By Combining Semantically Independent And Dependent Features
Author: Chun-Yu Yang, Hsin-Ho Yeh, Chu-Song Chen
Description: Study the aesthetic features, discover their semantic property on videos and then come up with more useful video based features such as motion space and motion direction entropy. It compares the assessing accuracy between two different semantic types of features and find that the semantic independent feature is more reliable from the results. However, the accuracy of the method for semantically dependent features has a huge margin of scope for improvement.

Title: Classification Of Video Media: Aesthetics
Authors: Pritesh S Patel , Madhura V Phatak , Ruhi A Patankar
Description: The aesthetics evaluation of video media can be used as useful clue to improve user satisfaction in many applications like search; broadcasting and recommendation. It automatically assess the quality of videos with a strong correlation self-defined /not standardized with human perception is a challenging task. However, there is no publically benchmarked datasets available till date and video shooting techniques.

Title: Towards Computational Models Of Visual Aesthetic Appeal Of Consumer Videos
Authors: Anush K. Moorthy, Pere Obrador, and Nuria Oliver

Description: It suggests controlled user study to collect unbiased ground truth about the aesthetic appeal of consumer videos, proposes frame-level and video-level features to characterize video's aesthetic appeal. There are 9 low-level features to characterize the aesthetic appeal of the videos. For features at the video level, evaluated various pooling strategies based on statistical measures. But, the drawbacks are: personalization techniques, lack of including universal vs. person-dependent and lack of assessing the influence of audio in aesthetic ratings.

Title: Photo And Video Quality Evaluation
Authors: Yiwen Luo and Xiaoou Tang
Description: It is based on professionalphotographytechniques, extract subject region from photo, formulates a number of high-level semantic features based on this subject and background division. Several important criteria used by photographers to improve photo quality rely on different treatment of the subject and the background. Only utilizes simple features, stresses more on subject background features does not take into consideration semantic features.

Title: Beauty Is Here: Evaluating Aesthetics In Videos Using Multimodal Features And Free Training Data
Authors: Yanran Wang, Qi Dai, Rui Feng, Yu-Gang
Description: Demonstrates a computational approach to automatically evaluate the aesthetics of videos, with particular emphasis on identifying beautiful scenes. Considers a large variety of features including not only low-level features, but also mid-level semantic attributes For a more broader task of evaluating video aesthetics, Image-based training being suitable for the scenario of the NHK Challenge may not hold.

Title: High Level Describable Attributes for Predicting Aesthetics
Authors: Sagnik Dhar, Vicente Ordonez, Tamara L Berg
Description: Demonstrates a simple, yet powerful method to automatically select high aesthetic quality images from large image collections. It introduced a method to produce query specific interestingness classifiers. The effect of social processes on the acquisition of knowledge, the finding of meaning and evaluation of meaning of relevant art works.

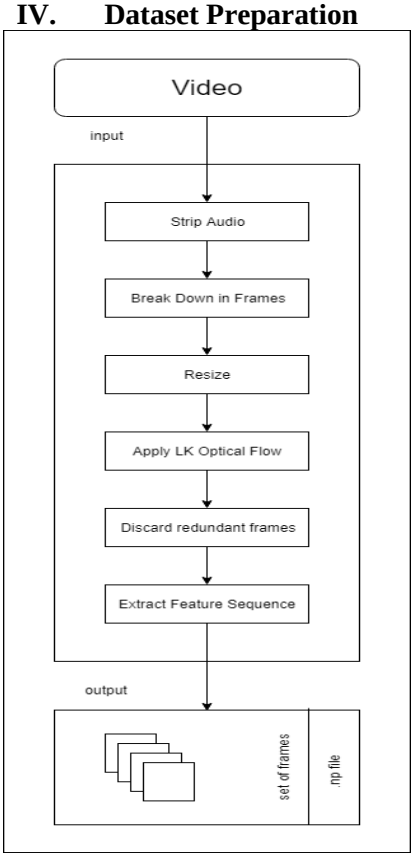
III. Dataset

YUP++ Dataset: This dataset includes samples from 20 scene categories of which 5 video categories are considered by us. They are Sky Clouds, Waterfall, Windmill, Railways and Highway. The duration of each video is 5 seconds approximately. Total number of videos is 60 videos per scene class i.e 1200 videos. Thus, 300 videos are considered by us.

CERTH-ITI-VAQ700: The duration of each video ranges from 1 minute to 6 minutes. Annotators evaluate and assign binary aesthetic ratings, 1 for Aesthetically Pleasing and 0 for Not Aesthetically Pleasing. Final score is the median score of individual annotators. This dataset exploits the features that are motivated by photography and cinematography rules. The total number of videos in this dataset is 700.

Self-Complied Dataset: This dataset has 50 normal paced videos classified as Aesthetically Pleasing. These videos are speed up 4x times using ffmpeg and classified as not appealing. The normal paced videos are then slowed down 0.25x times and classified as not pleasing. 50 videos incorporated with handshaking are classified as not pleasing. These videos are then smoothed out using Adobe Premier Pro and classified as Aesthetically Pleasing.

IV. Dataset Preparation



1. A video is given as input to the data preparation module.
2. The audio of this video is stripped off.
3. The video is then broken down into a series of frames.
4. The width and height of these frames are then resized to ensure uniformity of input to CNN.
5. The Lucas Kannade Optical Flow method is then applied to these resized frames.
6. Every video will be subsampled down to 40 frames. So a 41-frame video and a 500 frame video will both be reduced to 40 frames, with the 500-frame video essentially being fast-forwarded.
7. Features are extracted and stored in a file with the .np extension.

V. Features Considered

1. **Camera in Motion:** It alters the relationship between the subject and the camera frame, shaping the viewer's perspective of space and time and controlling the delivery of narrative information.
2. **Object in Motion: T**he camera frame is stable while the object within the frame moves .
3. **Handshaking: H**and-shaking occurs occasionally and has often been disturbing when the audience tries to concentrate in a video, thus making it significant to distinguish the high quality videos from low
4. **Motion Space:** It the space present in the frame for the portrayal of motion. It gives the camera the ability to track movements within a frame.
5. **Motion Entropy:** It gives us the amount of information contained in a picture. Two adjacent pictures in a video stream may have the same amount of information content but with the subject moved to a different area of the picture.

VI. Proposed System

1. User uploads a video with a length less than 15 seconds.
2. Motion key frames are identified using Lucas Kanade optical flow
3. Trained 3D CNN classifies if the video is aesthetically pleasing or not based on
   a) a measure of similarity between successive frames
   b) a measure of the diversity of motion directions
   c) a measure of the stability of the camera during the capturing process

d) a measure which can distinguish the difference between three categories of shots: focused shots, panorama shots and static shots.

4. Result is returned to the user

VII. System Architecture

Following diagram is our system's architecture

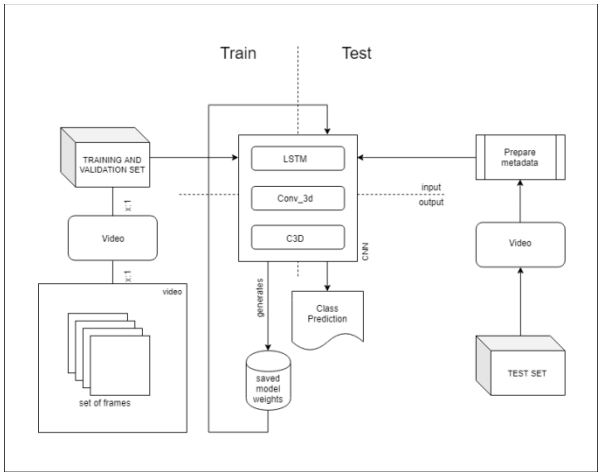


Fig 1: System architecture

We implement the proposed system using Keras in Python. The implementation is done using three models of deep learning that are c3d, Conv_3d and lstm.

C3D: It is a sequential model, consists of 6 layer groups. First two layer groups perform convolution and maxpooling in sequence. Third and fourth layer group perform two convolutions followed by maxpooling.
Fifth layer group performs two convolutions and a maxpooling operation, while also flattening the output
Final layer is a Fully Connected layer.

Conv_3D: Same as above but consists of only 3 layers.

LSTM: LSTM works on extracted feature sequence from a set of frames.
It is implemented as a sequential model, by adding an LSTM layer structure, followed by a fully connected group, with dropout.

Here, the Stochastic Gradient Optimizer is used. For the input layer of SGD, RELU activation Function is used whereas for the output layer, Sigmoid Activation function is used.

VIII. Methodology

Mathematical model:

M_1: REPOSITORY MODULE

INPUT = {V}
V: Array Of Video clips
OUTPUT = {V_1}
V_1: Uploading the video from V
F = { F_1, F_2}
F_1: Empty V
F_2: V_1 has missing attributes

M_2: PRE-PROCESSING MODULE

INPUT = {V_1}
V_1: Single video with valid attributes[V_1]
OUTPUT = {K}
K: Key Frames of V_1
F = {F_1}
F_1: Too many or very few key frames generated

M_3: FEATURE EXTRACTION USING HANDCRAFTED MECHANISM MODULE
INPUT = {K}
K : Key Frames Of V_1
OUTPUT = { C_M,C_S,O_M,O_S}
C_M: Camera is in motion
C_S: Camera is stationary
O_M: Object is in motion
O_S: Object is stationary
F = {F_1}
F_1: Object could not be distinctly identified

M_4: TRAINING DEEP EARNING FRAMEWORK MODULE
INPUT = {C_M, C_S , O_S, O_M}
C_M: Camera is in motion
C_S: Camera is stationary
O_M: Object is in motion
O_S: Object is stationary
OUTPUT = { MV_F,MV_S,MV_M,HS_P,HS_A,ME_H,ME_L }
MV_F: Motion Velocity Fast
MV_S: Motion Velocity Slow
MV_M: Motion Velocity Medium
HS_P: Hand Shaking Present
HS_A: Hand Shaking Absent
ME_H: Motion Entropy High
ME_L: Motion Entropy Low

IX. Results

For Generic Dataset of 100 videos:

Models	train_accuracy	train_loss	validation_accuracy	validation_loss
c3d	0.6920	0.4754	0.6924	0.4875
lstm	0.6931	0.5246	0.6931	0.5000
Conv3d	0.6917	0.5492	0.6931	0.5012

For Motion Based Dataset of 260 videos:

Models	train_accuracy	train_loss	validation_accuracy	validation_loss
c3d	0.6143	0.6877	0.5161	0.6828
lstm	0.5494	0.6930	0.5761	0.6931
Conv3d	0.5855	0.6877	0.6206	0.6836

Thus, on our dataset, the C3D model performs better than Conv_3d and LSTM model. c3d model works best as it takes images into consideration as opposed to LSTM which considers the extracted feature sequence of fixed length. With respect to c3d and Conv3d, c3d is a refined version of Conv3d architecture and thus provides a better result. The accuracy for all three models can be improved by: a) Standardizing the Video Aesthetic metrics, b) Improving the robustness of the Dataset. To successfully analyze the effect of motion on the aesthetic perception of videos, we need to delve deeper into areas of User Research, Digital Media, Visual Communication, and apply a Deep Learning Approach to a wide variety of input data.

X. Conclusion

We are presenting a system for classifying pleasing and non-pleasing videos. First, we are analyzing the perceptual differences between the two classes, and then we are then using DCNN as well as feature module to extract those differences. Using a diverse and difficult set of videos crawled from the web, we are training our classifier to use our modules. Thus, we can successfully capture the aesthetic essence of a video on highly accurate evaluating platform.

XI. Future Scope

Rating of videos according to its aesthetically pleasing nature, along with detailed suggestions and advice to improve the aesthetic quality can be done. Specific models could be designed/trained to make the video classification, task-specific. CNN could be trained to identify the keyframes of a given video. CNN could learn to extract salient motion features by itself.

References

[1]. Helmut Leder, Benno Belke, Andries Oeberst and Dorothee Augustin, "*A model of aesthetic appreciation and aesthetic judgments*" British Journal of Psychology (2004), 95, 489–508.

[2]. Yanran Wang, Qi Dai, Rui Feng, Yu-Gang Jiang, "*Beauty is Here: Evaluating Aesthetics in Videos Using Multimodal Features and Free Training Data*"School of Computer Science, Fudan University, Shanghai, China.

[3]. Ritendra Datta, Jia Li, and James Z. Wang, "*Algorithmic Inferencing Of Aesthetics & Emotion in NaturalImages:AnExpostion*" The Pennsylvania State University, University Park, PA16802,USA.

[4]. Sagnik Dhar,Vicente Ordonez, Tamara L Berg, "*High Level Describable Attributes for Predicting Aesthetics and Interestingness*" Stony Brook University,Stony Brook, NY 11794, USA.

[5]. Yan Ke, Xiaoou Tang, Feng Jing, "*The Design of High-Level Features for Photo Quality Assessment*" School of Computer Science, Carnegie Mellon; Microsoft Research Asia.

[6]. Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James. Z. Wang, "*Rating Image Aesthetics using Deep Learning*"

[7]. Yiwen Luo and Xiaoou Tang, "*Photo and Video Quality Evaluation: Focusing on the Subject*" Department of Information Engineering The Chinese University of Hong Kong, Hong Kong.

[8]. Ningrinla Marchang and Raja Datta. (2008). Collaborative techniques for intrusion detection in mobile ad-hoc networks. *elsevier*. 6 (n.d), p-508-523.

[9]. Hadi Otrok , Noman Mohammed, Lingyu Wang, Mourad Debbabi and Prabir Bhattacharya. (2008). A game-theoretic intrusion detection model for mobile adhoc networks. *elsevier*. 31 (n.d), p-708–721.

[10]. Sanjoy Ghatak, Key-Frame Extraction Using Threshold Technique

Mr. Shreyas Joshi. "Video Aesthetic Assessment Using Deep Learning." IOSR Journal of Computer Engineering (IOSR-JCE) 21.6 (2019): 12-17.