Like at First Sight: Understanding User Engagement with the World of Microvideos

1, 2
King's College London, UK
blah, bleh}@kcl.ac.uk

ABSTRACT

Several content-driven platforms have adopted the 'microvideo' format, a new form of short video that is constrained in duration, typically at most 5-10 seconds long. Microvideos are typically viewed through mobile apps, and are presented to viewers as a long list of videos that can be scrolled through. How should microvideo creators capture viewers' attention in the short attention span? Does quality of content matter? Or is there greater support for "rich gets richer" theories, which suggest a self-perpetuating phenomenon where content from users with large numbers of followers stands a greater chance of becoming popular? To the extent that quality matters, what aspect of the video – aesthetics or affect – is critical to ensuring popularity?

We examine these questions using a snapshot of nearly all videos uploaded to globally accessible channels on the microvideo platform Vine over an 8 week period. We find that although social factors do affect popularity, content quality becomes critical at the top end of the popularity scale. Furthermore, using the temporal aspects of video, we verify that decisions are made quickly, and that first impressions matter more, with the first seconds of the video typically being of higher quality and having a large effect on overall popularity. We demonstrate that despite being a video format, microvideos on Vine fall on a spectrum in between the more familiar user-generated images on sites like Flickr, and user-generated videos on sites like YouTube.

1. INTRODUCTION

In the last few years, we have seen the introduction of a new form of user-generated video, where severe restrictions are placed on the duration of the content. High-profile examples include Vine, which allows users to create videos up to 6 seconds long; Instagram, which introduced videos up to 15 seconds long; and Snapchat, whose videos are officially limited to 10 seconds. Although most user-generated video platforms have placed some form of limit on the duration or size of videos (e.g., YouTube had a 10 minute limit, which has since been softened to a 'default' limit of 15 mins[1]), the extremely short duration time limits of Vine etc has led to the coining of a new term: *micro videos*.

On the one hand, these time restrictions offer a way for the platforms to manage the size of the videos and thereby make the infrastructure management and content delivery more manageable. Thus, the time restrictions may be thought of as an artificial limit imposed on users. On the other hand, media commentators have argued that these restrictions could fundamentally change the way we communicate [5]. Similarly, David Pogue, writing in the Scientific American, conjectured that micro videos be seen as whole new form of expression that is more closely related to images than to videos [17]:

> A photograph is intended to capture a single moment, to present it for thoughtful examination. In the end, that's what a one- or six-second looping video does so well – it's just that it expands the scope of the still image ... Maybe the micro video is best considered an improvement on a still picture, not a downgrade from video.

This hypothesis raises the question of whether it is only *conceptually* convenient and interesting to model micro videos as enhanced images, or whether users also *engage* with micro videos in ways similar to images. The answer has strong implications for how micro video platforms are designed and used. For instance, Miriam - pl check rest of this para and make it sane from your perspective: an image-like medium may benefit from image-like editing tools such as filters, cropping, fitting etc, whereas a video-like medium would benefit from non-linear or multi-track editing features. More importantly, many micro video platforms have started relaxing the time restrictions (e.g., Instagram videos can now last up to 60 seconds, and Vine recently introduced an option where users can attach a longer video up to 140 seconds long that can be reached in a single click from the 6 minute micro video). If in fact users see value in micro videos as images with embedded live action, then such attempts may alter the very essence of the medium of expression.

This paper takes a first look at how users engage with micro videos, by using Vine, one of the earliest and most popular micro video platforms, as a case study. We start by crawling POP12K, a dataset of about 12,000 videos which have been deemed by Vine to be popular, and therefore, by definition, have engaged a large number of users. We complement this by collecting ALL120K, a dataset of nearly all ($\approx$ 120,000) videos that were uploaded to one of the 18 globally available channels on Vine.

For each video in our dataset, we derive XXX aesthetic and affect (sentiment)-related features (See Table XXX) both in the video and the audio tracks, and collect statistics of the number of times each Vine was looped through, reposted or 'revined', and liked by

[1] https://techcrunch.com/2010/12/09/youtube-time-limit-2/

different users. Basing our study on these features, we systematically examine the nature of user engagement in Vine on three levels: First, we compare the features of micro videos in POP12K with corresponding features of popular content on image-based and video-based user generated content platforms (Flickr and YouTube respectively) and find evidence suggesting that Vine falls on a spectrum between images and videos.

Next, we ask how these features vary over time in Vine videos, and discover a *primacy of the first second* phenomenon: the best or most salient parts of the video, whether in the aesthetic space or affect space, are more prevalent in the initial seconds of the micro video, suggesting that the authors are consciously or unconsciously treating Vines similar to images – in the initial part of the video, it is composed with aesthetics and affective quality in mind, resulting in a higher quality level; but quality declines

Finally, we take the counts of loops, reposts and likes as metrics of collective user engagement of the *consumers* of a video, and ask what factors affect these metrics. We develop a simple random forest classifier that is able to distinguish popular and unpopular items with high accuracy

Below this is construction material for elsewhere As with other user-generated content, users may engage either as *producers* of content or as *consumers*.

User engagement is both what users see and what they feel (cite). So we look at aesthetics as well as affect. Information comes in three channels: audio, video and sentiments. Vine engagement can be understood in terms of consumers and producers: what users produce and how they produce it.

We examine this question in three

the time restrictions shape user engagement with micro videos. To this end, we chose to examine Vine, one of the earliest and most popular micro video platforms, and also a platform with one of the shortest time duration restrictions (just over 6 seconds). We start by crawling POP12K, a dataset of about 12,000 videos which have been deemed by Vine to be popular, and therefore, have by definition, engaged a large number of users. We complement this by collecting ALL120K, a dataset of nearly all ($\approx$ 120,000) videos that were uploaded to one of the 18 globally available channels on vine.

To understand which aspects impact engagement, we look at both the video and audio signals in the micro video. We consider *aesthetics* – how well constructed the video is, as well as affect – the emotions the video can evoke. We ask if we can distinguish "popular" videos from the unpopular ones based on these different signals, and find that surprisingly, we can do a great

We first compare the popular micro videos in POP12K with popular user-generated images from Flickr and user-generated videos from YouTube, and find that micro videos appear to occupy a middle ground between images and videos. Then we ask whether and to what extent the

2. RELATED WORK

The work done in micro video analysis has been limited. Work by Miriam et.al [19] try to quantify and build on the notion of creativity. Work by [18] use textual sentiments to bring thousands of fiction novels to sentiment space and show that most novels follow 7 salient categories of stories. A paper by Nguyen et.al [16] collected more than 200 thousand micro videos from vine. A work done by Fontanini et.al [4] explore relevence of perceptual sentiments to popularity of a video, but the work done was on youtube viral videos, which have a much richer composition and structure. The problem of understanding what makes a visual media stick, has been a difficult one to solve. There are a few approaches to understand the aesthic and memorability aspects of an image [8] [3] [11]

[2] –> Likes does not propagate quickly through Flickr social network (so quality is important)

[20] –> no correlation between age and popularity; most photos gain most likes in the first week.

[22] –> performs quite well, predicting popularity through text annotating features.

3. BACKGROUND AND METHODOLOGY

3.1 Introduction to Vine

Vine[2] is a video sharing platform owned by Twitter, where users can create and share videos called "vines" up to 6 seconds long, a constraint which was designed with the goal of inspiring "creativity"[3]. Vine is primarily used as a mobile app, although a Web interface as well as an Xbox Live interface is available to view the videos. Users may create vines and upload them to the platform (typically through the Vine mobile app), view vines created by others, and follow other users whose vines they find interesting.

Users see vines created by others in one of several tabs: The home tab shows a personalised and social feed of videos created by those that they follow. The explore tab shows 20 feeds and channels. Eighteen of the channels are category-specific, such as 'Comedy', 'Music', 'Animals', 'Weird', 'Sports', 'Arts' etc; and appeal to different kinds of users based on their specific interests. Vines are assigned to specific channels by their creators. The remaining two channels are termed by Vine as 'popular-now' and 'on-the-rise', and are curated channels containing videos that have proven to be of of wider appeal to the entire vine population.

As with other modern content-sharing platforms, there is a social aspect: viewers can follow others whose vines they find interesting, 'like' or share interesting vines by 'revining' (i.e., reposting) the vines, and commenting on them. The numbers of likes, revines and comments are a measure of the popularity of a video. Uniquely, vine plays videos in a loop, going back to the beginning after reaching the end. Loops are repeated as long as the focus is on that video (e.g., video is active on the mobile phone screen if using the vine app, or mouse is on the video in the web version). Thus, letting a video play for more than one loop can be a sign of engagement. Therefore the platform also tracks and reports in real time the aggregate number loops across all users.

3.2 Dataset description

Dataset	Vines (total)	Loops (median)	Reposts (median)	Likes (median)
POP12K	11448	318566	2173	7544
UNPOP120K	122327	80	0	2

Table 1: Table

The data[4] used in this paper is summarised in Table 1, and were collected in two phases as described below:

3.2.1 Popular videos dataset

First, we collected $\approx$12,000 videos which have been marked by vine as 'popular', by tracking the 'popular-now' channel[5] over a

[2]http://vine.co

[3]http://blog.vine.co/post/55514427556/introducing-vine

[4]All datasets will be made available for non-commercial research.

[5]https://vine.co/popular-now

three week period in Dec 2015, and downloading all videos and associated metadata once every six hours, and removing any overlapping videos from the previous visit. The crawling period was chosen to ensure that consecutive crawls have an overlap of several videos, and this sufficed for all visits made to the website during the data collection period; thus the dataset we collected is a complete collection of all 'popular-now' vines during the 21 days under consideration.

Vine does not disclose the algorithm used to mark a vine as popular; yet we observe (see Table 1) orders of magnitude more loops, reposts and likes in the popular-now dataset than in the channel dataset. Thus we believe that the algorithm used by vine to select vines for the 'popular-now' channel is strongly affected by the numbers of loops/revines/likes. Note that the numbers of loops etc. were collected at the time of crawl, within a maximum of six hours of being posted on the 'popular-now' channel, which limits the possibility that the counts increased *as a result* of being featured on the popular-now channel. In the rest of the paper, we use the counts in the popular-now dataset to calibrate the definition of 'popular'. While there is a possibility that this is a biased proxy for global popularity, it nevertheless provides a baseline against which to compare all videos.

3.2.2 *All channel videos dataset*

In the second phase, we collected videos accessible from each of the 18 global vine channels or categoriesover a period of **8 weeks** from **Aug XX to Oct YY 2016**. Again, a crawling period of six hours was chosen for consecutive visits to the same channel, and the 100 most recent vines were fetched with each visit. As shown in Fig. 1, the vines returned has a significant overlap with vines fetched from the previous visit. Thus we believe that our dataset captures nearly all videos uploaded to vine and assigned to a channel. The only exception is the extremely popular comedy channel, for which we nearly always find more than 100 new videos (we only download the 100 most recent videos for the comedy channel). In total, this results in a dataset of ≈120,000 videos. We track the loop, revine and like counts over time, periodically updating each video's counts every three days until the end of the data collection effort.

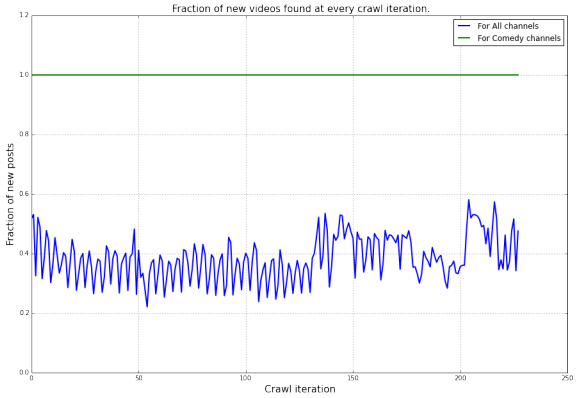

Figure 1: Fraction of new videos since last visit amongst the videos whose metadata was fetched by requesting for the 100 most recent videos from each channel is consistently less than 1, suggesting that the dataset contains nearly all videos from most channels. The only exception is the comedy channel, which consistently has more than 100 new videos (thus fraction of new videos is nearly always 1).

Note that while we obtain nearly all videos across the channels, our dataset does *not* capture *all* videos uploaded to vine – vine creators do not need to assign a video to a channel, and we do not discover any vines not in channels. We use channels to restrict ourselves to vines which get exposed to a reasonably global audience of those interested in a topic category, and therefore to vines that have a higher potential for garnering high like/revine/loop counts.

The popularity distribution of the whole dataset follows as expected a zipf distribution. The Fig. 1 shows the distribution of likes and repost counts of the collected videos on al og scale. Videos with 0 likes or reposts were given a marginal 1 like to avoid undefined logarithms.

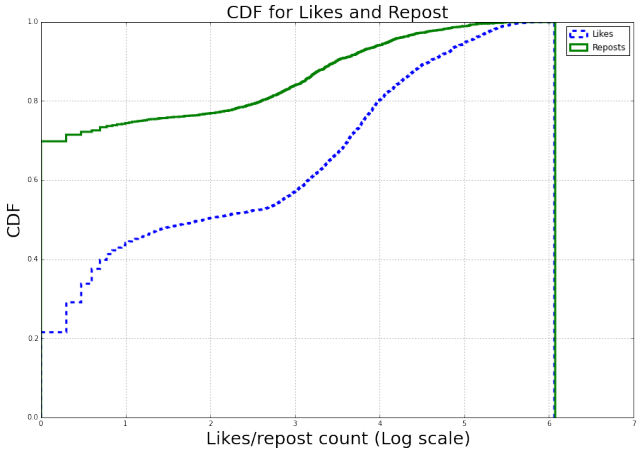

Figure 2: *CDF of Like count and Repost count.* **The values are normalized and on a Logarithmic scale. As expected from a long tail distribution of metrics like popularity, the dataset has a lot of videos with zero likes and reposts. Such videos are synthetically given one like and repost to avoid undefined values**

3.2.3 *Open datasets*

Over the course of this study, we also used several third party datasets, to corroborate different properties of vine videos, with datasets previously studied, ranked and evaluated by other research groups. This allowed us to do a comparative study about the postion of micro videos in the realm of social media modalities. For comparison with static web images, we use the popular MIR-Flickr dataset [7] which is an open source dataset of over 25,000 images crawled from the Flickr image service. The images were verified to be aesthetically pleasing. This was achieved by making sure that they were ranked high on the "interestingness"[6] rating. These images were used as a baseline for static web images in the comparison study. The other end of the spectrum of web media is Youtube. To compare, we used the dataset of above 400 viral youtube videos [9]. For our work with frame sentiments, we use the dataset provided by the Sentibank researchers [10] [1] for baselining the performance of our detector for detecting frame sentiments in micro videos.

4. VINE AS A NEW FORM OF EXPRESSION

As pointed out by previous work [19], micro-videos have been often called a "new form of expression" that goes beyond the traditional video medium. Vine constraints (short length, limited editing tools) offer an unprecedented number of possibilities for cre-

[6]https://www.flickr.com/explore/interesting

ative visual artists. Tech journalists [7] have indeed theorised that micro-videos cannot be even categorised as videos: their goal is to capture a single moment, thus making them ideally close to the photographic medium, or *"neither photo nor video but something in between, with artistic merits all its own"*.

In this Section, we understand the novelty introduced by the Vine platform in the media landscape using a computational approach. We consider 3 datasets of popular items from radically different platforms: Flickr (photography medium), Vine (micro-video medium) and Youtube (long video medium). We then compare these media over 3 dimensions: visual aesthetics, visual semantics and audio channel.

Datasets. We consider the following datasets for platform comparison. (1) *Photography medium:* we sample 1000 images from the MIR dataset (popular pictures from the Flickr stream) (2) *Micro-Videos medium:* we consider our POP12K dataset (popular Vines) (3) *Videos medium:* we use the 447 viral YouTube videos from the CMU dataset.

Visual Aesthetics Comparison. To compare the 3 platforms over the visual aesthetics dimension, we describe all items with the 18 computational aesthetics features from **SEC. BLA**. To obtain a description of the visual aesthetics trends for each medium, we aggregate the computed features at a platform-level. To do so, for each dataset, we compute the feature marginal probability distribution. This reflects the platforms' dominant visual aesthetics patterns (e.g. what are the common brightness values in YouTube videos? Or the distribution of Colorfulness in Vine?). To compare distributions across platforms (e.g. Rule of Thirds in Flickr VS Rule of Thirds in Vine), we then use symmetric Kullback-Lieber (KL) divergence, which reflects the distance of 2 probability distributions. The average of such KL divergence values tells us how far platforms are in the visual aesthetics feature space. We report the results of this analysis in Fig. 3: as expected, Vine videos show different visual aesthetic behaviour from both Flickr and Youtube, although more stylistically similar to long videos.
To understand why Vine is so different, we then look at the features with greater KL divergence across platforms. We notice that, in practice, Flickr aesthetic features reflect the behaviour of a "professional" medium, On the other side of the spectrum, micro-videos show patterns of less professional use, typical of the user-generated mobile-first Vine content. Youtube lies in the middle of such spectrum. More specifically, we notice the following (see Fig. 4). (1) *Colorfulness:* Flickr photos tend to have a higher color diversity, while Vine and Youtube tend to have less saturated colors. (2) *Exposure:* Flickr pictures tend to have a good balance between Left and Right pixel intensities, typical of high quality images; unlike Flickr, Youtube and Vine videos show unbalanced exposure. (3)) *Rule of Thirds:* Flickr pictures tend to deviate from the standard rule of thirds, typical of professional, artistic pictures. On the other hand, the moving images of Vine and Youtube tend to stick to the Rule of Thirds rule (4) *Sharpness* is probably one of the most important properties of high quality visual content. Due to its mobile-based nature, Vine videos tend to have almost no sharp pixels, while the professionally of Flickr photographers is clearly exposed by the higher percentage of sharp pixels.

Visual Semantics Comparison. To understand the semantic content depicted in Vine, Youtube and Flickr items, we use the deep learning-based object detectors from **CIT NEEDED**[]. For each visual item, we retain the labels of top-5 objects detected. We then aggregate such information at a platform level by computing the multinomial distribution of the detected objects for all Flickr images, Vine videos, and Youtube videos. Such distributions reflect the frequency of visual objects in typical popular videos of each platform (e.g. how often does a cat appear in a YouTube video?). Again, we then use symmetric KL divergence to compare object distributions across platforms. From Fig. **??**, we see that, in the object space, Youtube and Flickr are equally distant from Vine. By looking at element-wise differences across distributions, we then rank objects according to how different their frequencies are for the 3 platforms, and report results in Fig.5. Vine can be clearly distinguished from the other 2 media due to the higher presence of objects related to celebrations, fun, entertainment (*academic dress* ,*wig*,*tv*, ,*sunglasses*). Viral Youtube videos prefer popular subject such as kids (*diaper*), cars, and violent scenes (*punching*,*neck brace*). Finally, Flickr pictures can be distinguished for the presence of visual concepts typical of suggestive sceneries ((*lakeside*, (*seashore*.)

Audio Channel Comparison. The audio channel is as important as the visual channel for long videos. We want to understand here the role of audio in for Vine videos. For Youtube and Vine, we extract the audio features in **SEC. BLA**. Given the continuous nature of these features, we follow the same procedure used for aesthetic feature analysis (per-feature symmetric KL divergence). In Fig. 3, we see that in the audio space such media are far apart. This is due to the fact that, while the audio tracks of Youtube Videos are very diverse, and therefore follow almost-uniform distributions across different feature ranges, all audio tracks in Vine videos tend to have very similar patterns. In Vine, audio is mainly a weak complement to the visual counterpart: Vine videos can be fully consumed and understood without the audio channel, and they are often played in the mute mode. As a matter of fact, Vine videos tend to have few rythmical changes (low *Onset Rate*) and low *Roughness*. Also, overall, due to their less curated audios, Vine videos tend to be louder than Youtube videos (high *Energy*), as shown in Fig. 6.

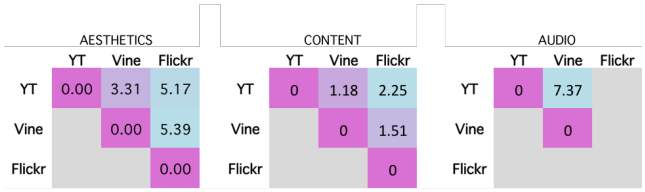

Figure 3: *Aggregated KL Divergences between platforms over all features for different feature groups.*

5. FEATURES

Over all the challege was to understand what makes a vine popular. And for that the there was a need to explore correlations of all the possible abstract midlevel features made available to us because of the rapid development in the fields of Deep machine learning. The main important contribution of machine learning is the ability of computationally extracting abstract higher level representations, which vaguely represent human perception.

5.1 Low level Aesthetic Features

: There are some well known computationally evaluatable aesthetic features which have been recognized as heuristics for good photography. Examples include Rule of thirds, Sharp Pixel proportion, Contrast, Simplicity, Left- Right symmetry etc [23]. The parameters basically compute perceptual features of an image based

[7] https://www.scientificamerican.com/article/why-micro-movies-so-popular-today/

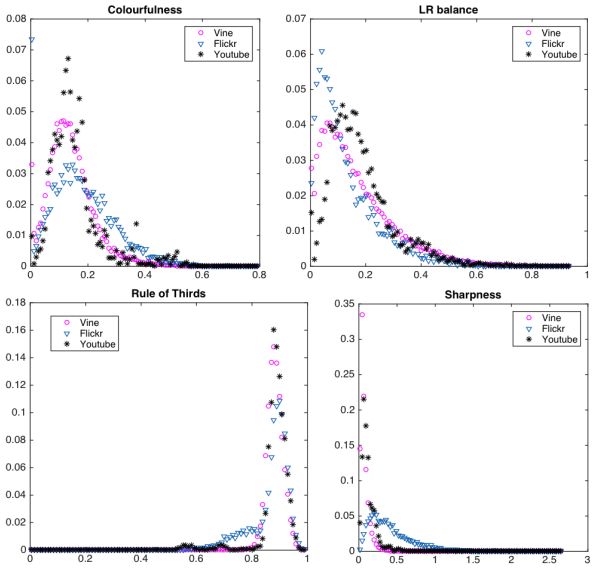

Figure 4: *Distributions of the most diverging aesthetic features across platforms.*

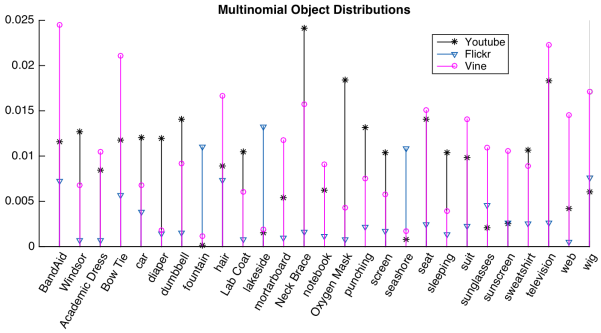

Figure 5: *Distributions of the most distant object occurrences across platforms.*

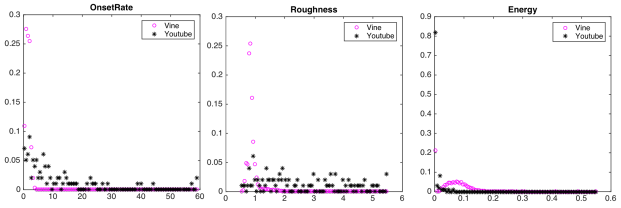

Figure 6: *Distributions of the most diverging audio features across platformss.*

on well know heurestic rules set by photographers. Some of the detailed references of the features we use are

5.1.1 Contrast

Contrast is basically dissmilarity between pixel(colour) values in a picture. It is a good aesthetic measure to understand how the photographer or creator of a visualc content has used the range of colour values to his advantage. This measure does not always reflect the aesthetic quality of an image, but with other features like sharp pixel proportion, can be a good approximation. For the sake of our study, we use Weber contrast, which is defined as

$$F_{weber} = \sum_{x=width} \sum_{y=height} \frac{I(x,y) - I_{average}}{I_{average}} \quad (1)$$

5.1.2 Simplicity

Simplicty of composition of a photograph is a distinguishible factor that directly correlates with professionalism of the creator [12]. We use simplicity definition as defined in [23] to calculate the ROI segment simplicity and Luo simplicity [15]

5.1.3 Rule of Thirds

This feature deals with compositional aspects of a photograph. Several papers including [23] study this feature and hence we use this as one of our aesthetic features. This feature basically calculates if the object of interest is placed in one of the imaginary intersection of lines drawn at approximate onethird of the horizontal and vertical postions. This is a well known aesthetic guideline for photographers.

5.1.4 Sharp Pixel Proportion

Out of focus or blurry photographs are generally not considered aesthetically pleasing. In this feature we measure the proportion of sharp pixels compared to total pixels. To do so we have to transform the image from intensity domain to frequency domain, and then count the total number of pixels which surpass the shapness criterion. We choose the criterion of sharpness in frequency domain to be 2 from [23]. The processing of the images was done using a tool called OpenIMAJ [6]

5.1.5 L-R Balance

Difference in intensity of pixels between two sections of an image is also a good measure of aesthetic quality. In non-ideal lighting conditions, images and videos tend to be over exposed in one part and correctly exposed in other. This is generally a sign of amature creator. To capture this we compare the distribution of intensities of pixels in the left and right side of the image. The distance between the two distributions is measured using Chi-squared distance.

5.1.6 Naturalness

This is a very heuristic property if an image that tries to gauge the degree of correspondence of images to the human perception. We first convert the image from RGB to HSL colour space which is proved to be closer to human perception of colours. We then group pixels using a heuristic rule that chooses pixels corresponding to natural objects like skin, grass, sky, water etc. This is done by choosing pixels which have L $\in$ [20 , 80] and S > 0.1. The final naturalness score is calulated by finding the weighted average of all the groups of pixels. [24]

5.1.7 Colourfulness

This is measure of an image's difference against a pure Gray image. It calculated as specified in [23]

5.2 Presence of Faces

One important aspect of micro videos is the presence of user as an actor in the video. When you look at viral vine videos, most videos seem to have a lead actor performing a skit. The hypothesis here was that vine has become a social media network, where actors have gained prominence and become a reason for popularity. So we did a small experiment, where we sample one image every second from all the videos collected. Then we calculate the percentage of frames across each video which contained at least one face in it.

We use the well tested Viola Jones dectector for frontal and profile face detection. [21]. When we plot the CDF of these percentages for popular against unpopular videos we see considerably higher population of popular videos to have high number of face image percentage 8.

5.3 Frame Sentiments

The crawled vine dataset was sampled and processed using the Multi Lingual sentibank detectors [10] which expresses visual sentiment of video feames on the scale of 1 to 5, 1 being negative and 5 being positive sentiments. To make use of this framework we randomly select about 12000 videos from the collection of 80k plus videos having both popular and unpopular vines. Then we sample these videos for frames twice every second. This creates a time series of video frames which could be now fed into the deep neural network for estimation of sentiments. These sentiments are basically vectors of length 12, (vine videos can be at most 6 seconds long) and there are 12000 such vectors. We can now do some statistical analysis on this N x 12 matrix

5.4 Audio Features

Along with aesthetic and perceptual sentiment features, audio is a big part of the Vine clip. We decided to use features that resonate more with the perceptual side of the analysis than the low level. We use 6 features that signify [19] perceptual attributes like loudness, rhythemical features, roughness etc. We extract these perceptual features using open source tools and incorporate them in the list of features we use for analysis [13] [14].

5.5 Social Features

To incorporate the effects of the social visibility for a post, the follower count and the past post count of the user who posts a particular video is used along with all the percieved features.

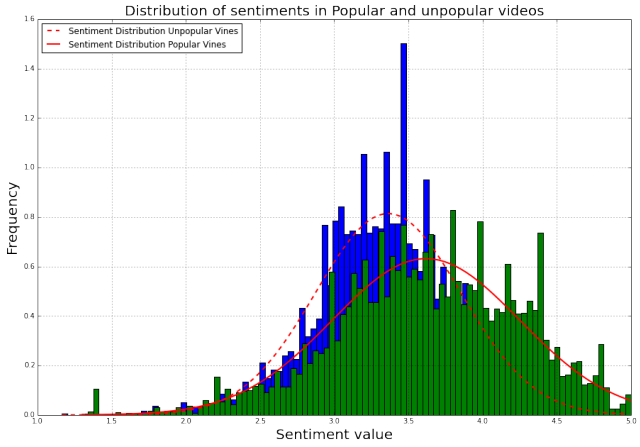

Figure 7: *Distribution of sentiment values for Popular and unpopular videos. The distributions follow a Gaussian like curve, but Popular videos tend to have more positive sentiments than Unpopular*

6. INSIGHTS FROM THE FEATURES

To understand the phenomenon of user engagement with the mico video content, we extract the above listed features from the sampled frames and extracted audio tracks of the videos. The motivation behind this excercise was to find peculiar discriminators amongst these features with respect to engagement. The following section describes some of the saliant differences between the high engagement and low engagement micro videos. To analyse the features, we sampled the videos twice every second and represented the whole 6 second long video as a series of 12 static frames. This sampling rate is not too low to miss any considerable frame transitions, neither is too high to include a lot of mid transition frames.

6.1 Faces engage us

For analysing this feature, we run the viola jones face detector for profile and frontal faces on every frame of the selected videos. Once sampled videos were processed for presence of faces, we analysed their prevalence seperately for the high engagement videos sampeld from the POP12K dataset and the low engagement videos sampled from the ALL120K dataset. The figure 8 shows the CDF of the percentage of face frames relative to total frames in a micro video. It is notable to see that highly engaging videos, tend to have a higher percentage of face frames. This shows a deeper behaviour of users, a tendency to be engaged to human faces.

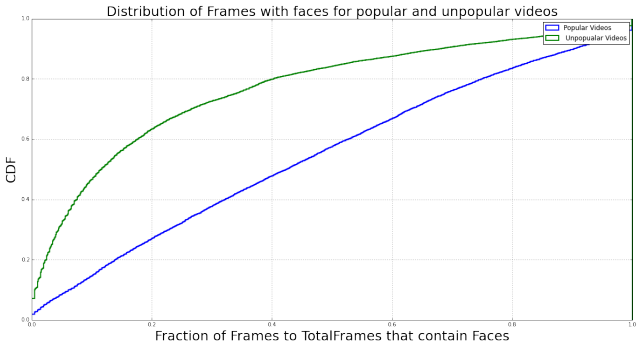

Figure 8: *CDF for popular and unpopular videos. The CDF signifies the cumulative distribution of percentages of face containing frames in a vine video. The observation here is popular videos tend to have higher percentages than unpopular videos*

6.2 Too short to make an impact ?

Image sentiments extracted using Sentibank framework are basically a good measure to understand the sentiment of the producer of the image. Because these detectors were trained on static images downloaded from flicker they do a good job of understanding some sort of an abstract sentiment from an image. We wanted to explore the time signatures of these abstract frame sentiments for micro videos. . To do this we processed all the sampled frames using the Sentibank deeplearning framework to estimate sentiment of each frame on a scale of 0 to 5. The detailed reasoning and description of this sentiment scale can be found in[10]. We now had a N x 12 matrix of sentiment time signatures for N videos. In our case N is 12,000 micro videos, 6000 from POP12K and 6000 from ALL120K dataset. To understand if there are any patterns in the sentiment transitions across the videos, the frame transition matrix of dimension 12000x12 was clusterd using K means clustering algorithm (cite). But to find the right amount of clusters we used the elbow method (cite). In the elbow method, k-means is iteratively run over the complete dataset for different values of K ranging from 1 to a resonably high maximum. We run till K = 10. At each step the sum of squred distance is calculated for every cluster between the cluster centroid and all the other vectors which are member of

Feature Name	Dimensions	Description
Mean Sentiment	1	Mean of sentiments detected using sentibank [**?**]
Contrast	3	Frame contrast calculated using Webber, color and RMS techniques
Simplicity	2	Image simplicity calculated by two methods [23]
Naturalness	1	A measure of "Naturalness" of a frame
Colourfulness	1	A measure of colourfullness that describes the deviation from a pure gray image
Hue Stats	2	Hue mean and variance which signifies the range of pure colours present in the image
LR balance	1	The Chi squared distance between the histogram of Left and Right side of image pixels.
Object Saliency	2	Measure of prominance given to salient objects. Includes Rule of thirds and ROI proportion
Image brightness	3	Features signify brightness of the image. Includes average brightness, saturation and saturation variance
Image sharpness	2	Features signify how sharp an image is. Includes sharpness variance and sharp pixel proportion
Audio Rhytmical Features	2	Onset rate and zero crossing rate which talks about rhythmic component of track [13]
Loudness	2	Overall energy and average short time energy which signifies loudness of the track [13]
Mode	1	Musical mode of the audio tract (major or minor). [13]
Roughness	2	measure of dissonance values between all peak pairs in the track [13]
Face Percentage	1	Percetage of frames in a video, which have been tested positive for atleast one face [21]
Social Features	2	Number of followers and past number of posts uploaded

Table 2: Dimensionality and description of features used for training the Classifier

the cluster.

$$SSD = \sum_{i=1}^{K=10} \sum_{X \in N} dist(X, c_i)^2 \quad (2)$$

We then plot the array of SSDs and look for and change in rate of decrease of SSDs, also called as the elbow point, on the graph. The value of K for which this point occurs has the most tight grouping of points in the 12 dimentional space when clustered in K clusters. From Fig 9 , this point is found to be at K = 4 in our case.

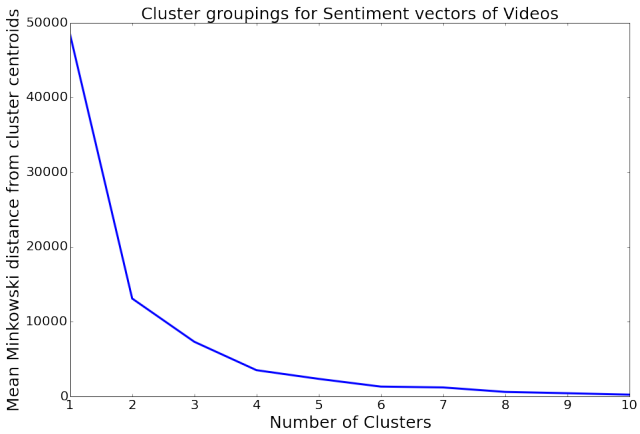

Figure 9: *The Sentiment transition matrix was analysed for existence of clusters, using the elbow points method for mean Euclidean distance (Minkowski distance for p = 2). We found that the best grouping exists at K=4*

Using this elbow point, we cluster the Matrix and look for salient trends in transition, by plotting the values of centroids of each cluster. The Fig 10 shows that despite being seperated in the sentiment space, the values of sentiments remain constant for a given cluster. The paper does not explore the fundamental reason behind this, but it is a peculiar behaviour. One possible hypothesis behind this could be that 6 seconds is too short a period for the creators of Micro videos, to project any kind of percievable and impactful change in frame sentiments.

6.3 First impression lasts

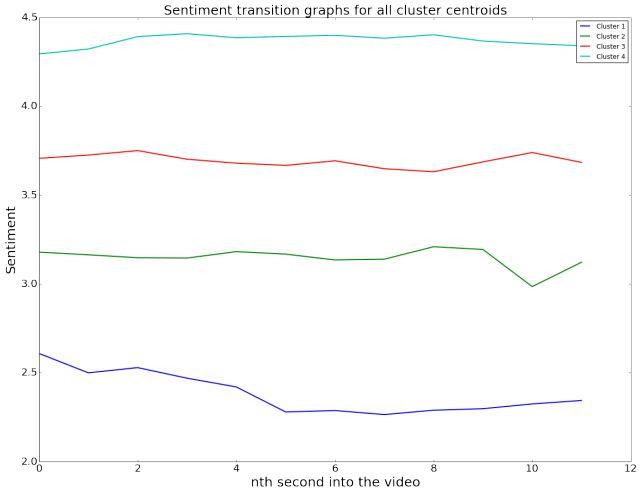

Figure 10: *Plot of frame sentiment vectors, of the 4 centroids of the clusters found. The vine videos tend to have a constant sentiment structure at 4 distinct sentiment levels*

The whole idea of a user engaging with a micro video on Vine, involves the process of scrolling past videos and then keeping a video in focus so as to trigger the auto looping mechanism described in introduction. This whole user experience makes engagement highly influenced by how far the user plays the video. To understand the effect of frame sentiments in different parts if vidoe over this user behavious, we plotted a simple histogram of the frequency of occance of maximum or minumum frame sentiments against the thirds of video. A trend that was evident from this was that the probability of coming across the most positive or the most negative sentiment in a micro video is the highest in the first one third of the video and progresively decays (Fig 11

We repeated a similar process for aesthetic features. There too, we found a similar trend of the first one third of video having the best overall aesthetic quality (12

These two trends say something about the producer behaviour. How should I close this insight? What should I draw any conclusion from this or let the classifier experiment talk about it ?

6.4 Quality matters, but not that much

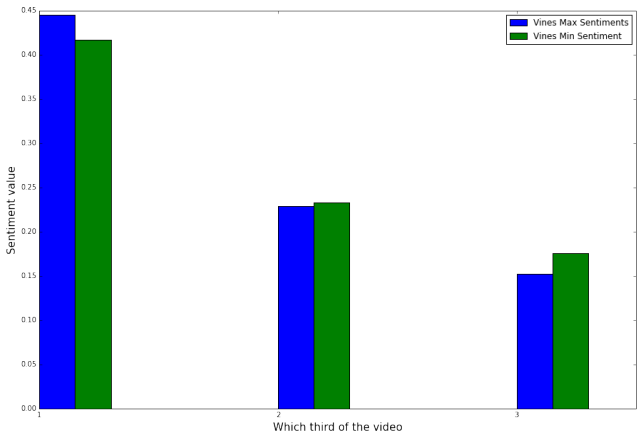

Figure 11: *Frequency of occurance of maximum or minimum sentiment. For this graph, the video is considered in thirds, and the frequency of occurance of both max and min frame sentiments is plotted.*

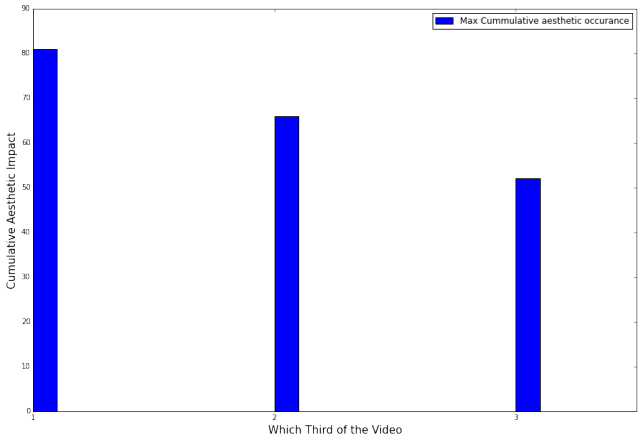

Figure 12: *Plot of cumulative aesthetic impact of each third of the video. For this plot the videos were sampled at one frame a second, and aesthetic features were calculated at every second. Finally the features were*

Aesthetics are important part of any social media analysis. They are going to play a part in garnering user engaagement. But how much? To comment about this question, we needed a comparision of overall aesthetics of micro videos with some sort of a gold standard for highy aesthetic images. Hence we compare the aesthetic features of frames sampled from both high engagement micro videos and low engagement videos, with images taken from the dataset sourced from photo.net [3]. These images were rated for aesthetic appeal for the research involved and these ratings are also used while selecting the images for our comparison. We only choose images with median ratings of 6 or above on aesthetic scale of 0 to 7. The table 3 shows the comparison of means and medians of several of these aesthetic features compared to the highly aesthetic dataset. From the table it is quite evident that Aesthetics in micro videos matter, but they are no way close to what one can call aesthetically pleasing frames.

7. DESIGNING A CLASSIFIER

Table 3: List of Aesthetic parameters computed for highly rated aesthetic images, Popular videos and unpopular videos. Most parameters have no bias towards either popular or unpopular videos

Parameter	Aesthetic Images		Popular Vines		Unpopular Vines	
_	Mean	Median	Mean	Median	Mean	Median
Color Contrast	51.05	30.22	29.88	16.43	20.23	8.83
Intensity Balance	0.11	0.08	0.16	0.13	0.17	0.14
Luo Simplicity	0.009	0.005	0.013	0.012	0.015	0.014
Sharp pixel proportion	0.103	0.098	0.090	0.085	0.089	0.081
Image Saturation	0.943	0.974	0.672	0.678	0.615	0.646
Avg. Brightness	0.148	0.141	0.137	0.130	0.139	0.124
Rule of Thirds	0.879	0.899	0.883	0.883	0.878	0.882
ROI Proportion	0.316	0.089	0.175	0.112	0.165	0.110

To understand the processes involve in user engegement with a micro video, we designed an experinmen where in we use machine learning methods, to understand influence of the features used for training, on the engagement classification process. By doing so we would be able to understand the relevance of different kinds of features on the overall engagement potential of a micro video and gain some empirical evidence about the impact contribution by each Aesthetic, Audio, contextual and social features. Further to validate our hypothesis about differential impact of the initial seconds on the engagement against the rest of the video, we train two classifiers where one uses the features across the complete video, and other only the first 2 seconds. Comparing the performances of both, we were able to conclusively propose the validity of our hypothesis.

Due to the exhastive nature of our data, the computational resources in terms of CPU usage and time would be unmanagable, so we sampled 18,000 videos from our datasets to work with. Out of thes 18,000 6,000 videos were sampled from the POP12K gold standard dataset which contain all high engagement videos. Addional 12,00 videos were sampled from the broadly sampled UNPOP120K dataset. These videos mostly fall towards the lower end of engagement. These videos were then sampled for individual frames every second and processed to extract the 28 dimentional vector of all the contextual, perceptual , aesthetic features. Audio features were seperately extracted from the audio track and social fearures from the post metadata. To get a better understanding of the dimensionality and ordering of the features refer to Table 2. The classifier is binary, so it only tells us if a video is engaging or not for a given definition of "Engaging". Using this as a criteria, we vary our definition of "Engaging" by changing the threshold of Loop count, at which a video is labelled as engaging. This threshold is varied from median loop count of ALL120K dataset all the way till 1.5 times Median of POP12K dataset. So in a nutshell, we make our definition of what classifies as engaging, more and more selective. The classifier is trained using cross validated dataset, which is devided in 80% for training and 20% for testing. The training process is interated across the range of engagement threshold criteria , 30 times, and in each iteration the performance and the impact of individual features is noted. The Fig 13 shows the variation of the impact of social features, compared with all the other content related features across all the iterations. It is important to note, that the impact of content related features, become more decisive as we make our engagement criteria more selective.

Now to further verify the theory about the importance of first seconds in vine, we trained the whole classifier setup over the same dataset. The major difference this time was that the features were now extracted from the frames sampled from the first two seconds of the vine. HERE AFTER WE SHOW THAT THE PERFORMACE IS THE SAME OR DEGRADED MARGINALLY

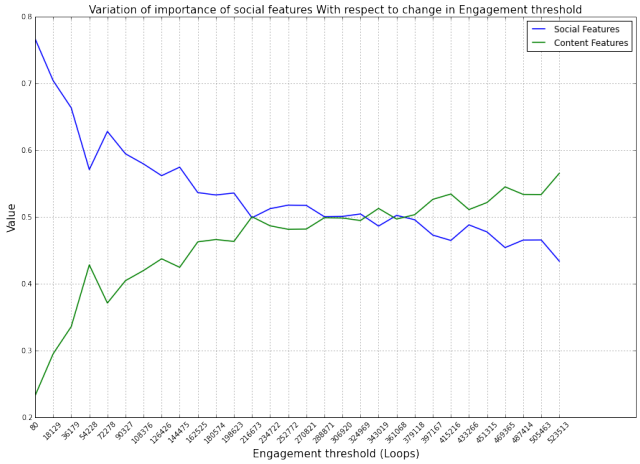

Figure 13: *A plot of contribution of social features against all the perceptual features combined. The influence is calculated by training the classifier and looking at the coeeficients of the final function. The values are generated by interating the training and testing process for different thresholds for the definition of a popular post. The x axis signifies the threshold value of Likes at which a video is labelled to be popular. It starts from the median of the UNPOP120K and ends at 1.5 times the median of POP12K.*

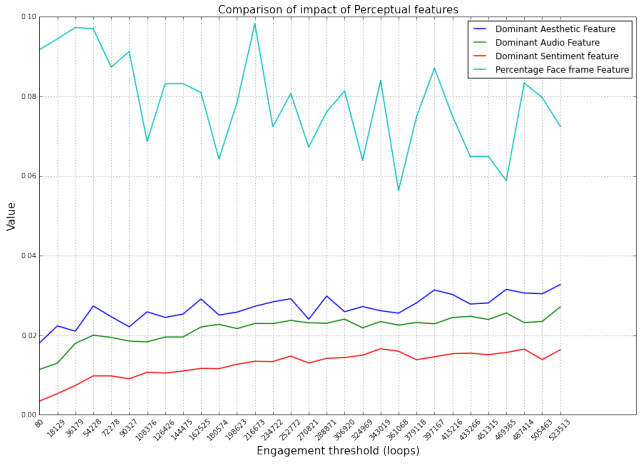

Figure 14: *The plot shows change in impact of individual track features as we become more selective. The most prominent features among the content features is the presence of face.*

8. DISCUSSION AND CONCLUSIONS

Microvideos are no longer micro – Instagram has extended from 15 – 60 seconds; Vine has extended to 140 seconds, similar to parent company Twitter's restriction of 140 characters etc. As the restrictions become less stringent, it would be interesting to study whether the essence of microvideos will continue to be the same, because the modes of content production are the same (mobile phone-based apps), or whether microvideos start behaving more like videos.

We find support for [17] who hypothesizes that microvideos bridge images and videos:

> A photograph is intended to capture a single moment, to present it for thoughtful examination. In the end, that's what a one- or six-second looping video does so well-it's just that it expands the scope of the still image . . . Maybe the micro video is best considered an improvement on a still picture, not a downgrade from video.

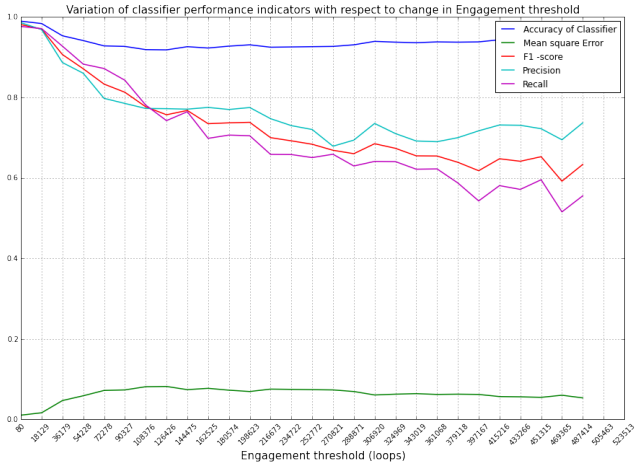

Figure 15: *The plot shows varying values of Precision, Recall, Accuracy and F1-score across the classifier traning iterations*

9. REFERENCES

[1] BORTH, D., JI, R., CHEN, T., BREUEL, T., AND CHANG, S.-F. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia* (New York, NY, USA, 2013), MM '13, ACM, pp. 223–232.

[2] CHA, M., MISLOVE, A., AND GUMMADI, K. P. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th International Conference on World Wide Web* (New York, NY, USA, 2009), WWW '09, ACM, pp. 721–730.

[3] DATTA, R., LI, J., AND WANG, J. Z. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *2008 15th IEEE International Conference on Image Processing* (2008), IEEE, pp. 105–108.

[4] FONTANINI, G., BERTINI, M., AND DEL BIMBO, A. Web video popularity prediction using sentiment and content visual features. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval* (2016), ACM, pp. 289–292.

[5] GROSSMAN, D. Can micro video change how we communicate? BBC Newsnight, Sep 2013.

[6] HARE, J. S., SAMANGOOEI, S., AND DUPPLAW, D. P. Openimaj and imageterrier: Java libraries and tools for scalable multimedia analysis and indexing of images. In *Proceedings of the 19th ACM international conference on Multimedia* (New York, NY, USA, 2011), MM '11, ACM, pp. 691–694.

[7] HUISKES, M. J., AND LEW, M. S. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval* (New York, NY, USA, 2008), ACM.

[8] ISOLA, P., XIAO, J., TORRALBA, A., AND OLIVA, A. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2011), pp. 145–152.

[9] JIANG, L., MIAO, Y., YANG, Y., LAN, Z., AND HAUPTMANN, A. G. Viral video style: A closer look at viral videos on youtube. In *Proceedings of International Conference on Multimedia Retrieval* (New York, NY, USA, 2014), ICMR '14, ACM, pp. 193:193–193:200.

[10] JOU, B., CHEN, T., PAPPAS, N., REDI, M., TOPKARA, M., AND CHANG, S.-F. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM international conference on Multimedia* (2015), ACM, pp. 159–168.

[11] KALAYEH, M. M., SEIFU, M., LALANNE, W., AND SHAH, M. How to take a good selfie? In *Proceedings of the 23rd ACM International Conference on Multimedia* (New York, NY, USA, 2015), MM '15, ACM, pp. 923–926.

[12] KE, Y., TANG, X., AND JING, F. The design of high-level features for photo quality assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (2006), vol. 1, IEEE, pp. 419–426.

[13] LARTILLOT, O., AND TOIVIAINEN, P. A matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects* (2007), pp. 237–244.

[14] LAURIER, C., LARTILLOT, O., EEROLA, T., AND TOIVIAINEN, P. Exploring relationships between audio features and emotion in music.

[15] LUO, Y., AND TANG, X. Photo and video quality evaluation: Focusing on the subject. In *European Conference on Computer Vision* (2008), Springer, pp. 386–399.

[16] NGUYEN, P. X., ROGEZ, G., FOWLKES, C., AND RAMAMNAN, D. The open world of micro-videos. *arXiv preprint arXiv:1603.09439* (2016).

[17] POGUE, D. Why are micro movies so popular these days? *Scientific American* (May 2013).

[18] REAGAN, A. J., MITCHELL, L., KILEY, D., DANFORTH, C. M., AND DODDS, P. S. The emotional arcs of stories are dominated by six basic shapes. *arXiv preprint arXiv:1606.07772* (2016).

[19] REDI, M., O'HARE, N., SCHIFANELLA, R., TREVISIOL, M., AND JAIMES, A. 6 seconds of sound and vision: Creativity in micro-videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 4272–4279.

[20] VALAFAR, M., REJAIE, R., AND WILLINGER, W. Beyond friendship graphs: A study of user interactions in flickr. In *Proceedings of the 2Nd ACM Workshop on Online Social Networks* (New York, NY, USA, 2009), WOSN '09, ACM, pp. 25–30.

[21] VIOLA, P., AND JONES, M. J. Robust real-time face detection. *International journal of computer vision 57*, 2 (2004), 137–154.

[22] YAMASAKI, T., SANO, S., AND AIZAWA, K. Social popularity score: Predicting numbers of views, comments, and favorites of social photos using only annotations. In *Proceedings of the First International Workshop on Internet-Scale Multimedia Management* (New York, NY, USA, 2014), WISMM '14, ACM, pp. 3–8.

[23] YEH, C.-H., HO, Y.-C., BARSKY, B. A., AND OUHYOUNG, M. Personalized photograph ranking and selection system. In *Proceedings of the 18th ACM international conference on Multimedia* (2010), ACM, pp. 211–220.

[24] ZHONG, C., KARAMSHUK, D., AND SASTRY, N. Predicting pinterest: Automating a distributed human computation. In *Proceedings of the 24th International Conference on World Wide Web* (New York, NY, USA, 2015), WWW '15, ACM, pp. 1417–1426.