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

Abstract

Several content-driven platforms have adopted the 'micro video' format, a new form of short video that is constrained in duration, typically at most 5-10 seconds long. Micro videos are typically viewed through mobile apps, and are presented to viewers as a long list of videos that can be scrolled through. How should micro video creators capture viewers' attention in the short attention span? Does quality of content matter? Or do social effects predominate, giving content from users with large numbers of followers a greater chance of becoming popular? To the extent that quality matters, what aspect of the video – aesthetics or affect – is critical to ensuring user engagement?

We examine these questions using a snapshot of nearly all videos uploaded to globally accessible channels on the micro video platform Vine over an 8 week period. We find that although social factors do affect engagement, content quality becomes equally important at the top end of the engagement 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 user engagement. We demonstrate that despite being a video format, micro videos 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.

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 allowed users to create videos up to 6.5 seconds long; Instagram, which introduced videos up to 15 seconds duration; and Snapchat, whose videos are officially limited to 10 seconds and are deleted after 24 hours. 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*.

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

Some media commentators have argued that the restrictions imposed by the micro video format could fundamentally change the way we communicate (Grossman 2013). Indeed, it has been argued that in its short lifetime, Vine has had a significant cultural impact far beyond its user base, generating several widely shared memes in its short lifetime[2]. At the same time, as the format is still very new, virtually all major micro video platforms are experimenting with the format, making significant changes in the last year. For instance, Instagram extended the limit from 15 seconds to 1 minute[3]. Snapchat has created a new wearable, called "Spectacles": sunglasses fitted with cameras that allow users to create and post 10-second long videos on the Snapchat platform[4]. Vine is undergoing a major overhaul – Twitter recently said it would close down the Vine website and community. The new version of the Vine app retains the 6.5 second video format, but the videos will be published directly on Twitter's feed and thus more closely integrated with its social network[5].

Our goal is to understand and shed light on the salient features of this evolving medium: How do micro videos differ from other user generated videos (e.g., on YouTube)? How does the strict time limit impact video quality, and user engagement (both as creators and consumers) with such videos? And as a corollary, do longer time limits change quality or user engagement? What are the relative roles of social and content quality factors in driving engagement and popularity – Is Twitter's move with Vine well motivated?

We answer these questions from an empirical perspective, using a dataset of nearly all ($\approx 120,000$) Vine videos that were uploaded to one of the 18 globally available channels on Vine during an 8 week period. We complement these with other datasets including a curated dataset (*POP12K*) of 12,000 popular Vine videos, as well as samples from other platforms – Instagram, Flickr and YouTube[6].

Using these datasets, we systematically examine the nature of user engagement in Vine on three levels. First, we

[2] http://www.theverge.com/2016/10/28/13456208/why-vine-died-twitter-shutdown

[3] http://www.theverge.com/tech/2016/3/29/11325294/instagram-video-60-seconds

[4] https://www.spectacles.com/

[5] http://www.theverge.com/2017/1/5/14175670/vine-shutting-down-rebrand-download-archive

[6] Flickr and YouTube data are publicly shared. On acceptance, our newly crawled data will also be shared for non commercial research

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 for the first time quantitative evidence that lends support to David Pogue's theory of micro videos as a whole new form of expression falling on a spectrum between images and videos (Pogue 2013).

Next, we take the three metrics of popularity we collect – counts of loops, reposts and likes – as quantification of the *collective* user engagement of the consumers of a video, and ask to what extent the content- and social network-related features affect these metrics. To answer this question, we adopt a novel methodology. We train a random forest classifier that, given a threshold for a metric of popularity, is able to able to distinguish items on either side of the threshold into popular and unpopular classes with high accuracy, precision and recall, using the features we have identified. The relative importance of different features then gives an indication of the extent to which those features affect the metric under consideration. We progressively consider higher and higher threshold values for videos to qualify as popular or engaging, and thereby identify trends and changes of relative importance of different features. Interestingly, we find that as the threshold for popularity becomes more and more stringent, features that represent quality of the content become collectively as important as social features such as the number of followers. Echoing an effect also observed in Instagram photos (Bakhshi and others 2014), we find that presence of faces significantly increases engagement, and is the most important content-related factor.

Finally, we ask how these features vary over time in micro-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 micro-videos similar to images – in the initial part, the video is composed with aesthetics and affective quality in mind, resulting in a higher quality level; but quality declines as the video plays over time. Furthermore, echoing the primacy of the first second phenomenon, we find that the quality of the first seconds of the video are as effective as the quality of the whole video in predicting popularity/engagement. Fig. 1 shows examples of these effects through two videos in our dataset of popular videos. In both videos, we observe content quality deteriorate over time, illustrating the primacy of the first seconds.

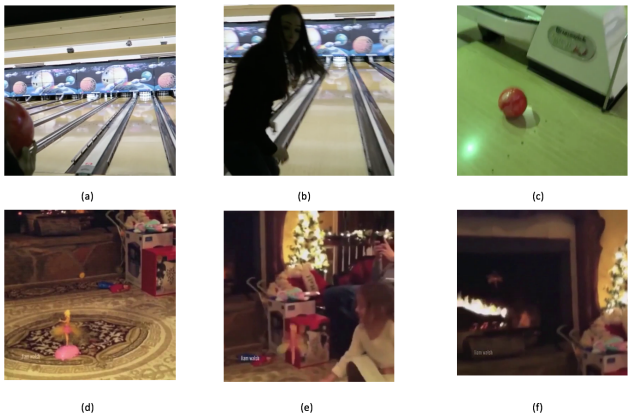


Figure 1: *Vine Samples from first, second and thirds one thirds of the video. Images (a) , (b) and (c) show a progressive drop in brightness and sharpness due to shaky camera. Images (d) (e) and (f) shows a progressive drop in contrast.*

Related Work

Our paper closely relates to those works in machine vision that infer intangible properties of images and videos. While computer vision frameworks typically focus on analysing image semantics using deep neural networks (Krizhevsky and others 2012), researchers have started exploring concepts beyond semantics, such as image memorability (Isola et al. 2011), emotions (Machajdik and Hanbury 2010), and, more broadly, pictorial aesthetics (Datta and others 2008; Luo and Tang 2008; et.al 2015). This work specifically focuses on online visual content collected from social media. Researchers have shown that, by leveraging social media data in combination with vision techniques, systems can estimate visual creativity (Redi and Others 2014), sentiment (Wang and others 2015; Jou and others 2015) and sarcasm (Schifanella and others 2016).

More specifically, our work closely relates to research that combines social media studies and computer vision to analyse popularity and diffusion for social media posts: for example, Zhong *et al.* were able to predict the number of post "re-pins" given the visual preferences of a Pinterest user (Zhong and others 2015); recent work (Mazloom and others 2016) has also used multimodal features to predict the popularity of brand-related social media posts. Different from these works which focus on prediction, this paper looks at understanding user engagement.

Media popularity prediction studies generally focus on non-visual features. For example, (Yamasaki and others 2014) used textual annotations to predict various popularity metrics of social photos. Social metrics such as early views (Pinto and others 2013) or latent social factors (Nwana and others 2013) have also been used to effectively estimate video popularity. However, the fact that many popular media items may not depend on the social network (Cha and others 2009) suggests that intrinsic media quality is an important factor for diffusion, engagement and popularity, which we explore in this paper.

Recent work in the field has explored the importance of visual content in analysing popularity: (Totti and others 2014) analysed the visual attributes impacting image diffusion, and (Schifanella and others 2015) studied relations between image quality and popularity in online photo sharing platforms. Bakhshi et al (Bakhshi and others 2014) showed that pictures with faces tend to be more popular than others. Similar to our paper, researchers have used computer vision techniques to estimate image popularity in Flickr (Khosla and others 2014). Moreover, a work done by Fontanini et.al (Fontanini and others 2016) explore relevance of perceptual sentiments to popularity of a video. Unlike these works,

we explore content features to fully understand user engagement and popularity in micro videos, a new form of expression radically different from both the photo medium and the video medium. We motivate our study by providing quantitative evidence for such radical novelty introduced by Vine videos, running a cross-platform comparison study based on audiovisual features.

Micro videos are relatively new, so work specifically on micro video analysis has been limited. Redi *et al.* (Redi and Others 2014) quantify and build on the notion of creativity in micro-videos. A large dataset of 200K Vine videos was collected by Nguyen et.al (Nguyen et al. 2016), focusing on analysis of tags. Closest to our work is Chen *et al.* (Chen and others 2016) who use multimodal features to predict popularity in micro videos. However, although we use popularity prediction as an intermediate tool, our focus is on understanding impact and importance of different features in determining popularity or engagement. To this end, we introduce a novel methodology that allows understanding up to which point social features are prominent over content features. Additionally, we demonstrate the "immediacy" of engagement with micro videos by showing that the content from the first two seconds of the video is just as good at predicting popularity as the entire content . Collectively, these results allow us to characterise Vine as a new medium of expression, different from previous work.

Introduction to Datasets

For this work, we wanted to have a diverse dataset from at least the most popular micro-video sources. We chose Vine[7]. Vine was launched in 2012 ,where users are constraint by duration of the video. Users create videos to a maximum length of 6 seconds using the mobile app. The videos are posted on user's profile which can be followed and shared by other users. Being the pioneer in micro-video format and being a platform solely designed for micro-videos, we stress most of our work on videos sampled from Vine dataset.

Dataset description

Dataset	Vines (total)	Loops (median)	Reposts (median)	Likes (median)
POP12K	11448	318566	2173	7544
ALL120K	122327	80	0	2

Table 1: Summary characteristics of datasets used

The data used in this paper is summarised in Table 1, and was collected in two phases as described below:

Popular videos dataset.

First, we collected $\approx$ 12,000 videos which have been marked by Vine as 'popular', by tracking the 'popular-now' channel[8] over a 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 non-popular 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 'high engagement'. While there is a possibility that this is a biased proxy for global engagement, it nevertheless provides a baseline against which to compare all videos.

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 16 to Oct 12 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. The number 100 was a result of an API limit from Vine. 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 $\approx$120,000 videos. We track loop, revine and like counts over time, periodically updating each video's counts every three days until the end of data collection. At the last tracking cycle, we have metadata for each post for 3 weeks post upload.

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. However, due to the Vine platform structure, vines that are not in channels have near-zero probability to get seen by other users apart from the followers. We use channels to restrict ourselves to vines which have a chance to 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 engagement.

Feature Descriptions

In order to fully understand how micro-videos engage users, we characterise the content of videos using computer vision and computational aesthetics techniques and extract a number of features(Table 2), which can be divided into the following categories:

[7] http://vine.co

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

Image quality features These features are mostly taken from computational aesthetics literature, and have been recognised as heuristics for good photography. Prior work (Zhong and others 2015) has identified a set of image quality features that robustly predict user interest in images. We adapt these to videos by computing the features on images taken at regular intervals from the video under consideration, and use the values to understand intrinsic quality of Vine videos. We use a combination of low-level features such as contrast, colourfulness, hue saturation, L-R balance, brightness and sharp pixel proportion, together more high-level features such as simplicity, naturalness of the image, and adherence to the "rule of thirds" heuristic.

Audio features Following previous work on micro videos (Redi and Others 2014), we use audio features known to have an impact on emotion and reception. Using open source tools(Lartillot and Toiviainen 2007; Laurier et al. 2009), we measure *loudness* (overall volume of the sound track), the *mode* (major or minor key), *roughness* (dissonance in the sound track), and *rythmical* features describing abrupt rhythmical changes in the audio signal.

Higher Level features Affect (emotions experienced) is well known to strongly impact on user engagement (O'Brien and Toms 2008; Leung 2009). To understand the sentiment conveyed by the video frames, we use the Multi Lingual Sentiment Ontology detectors (Jou and others 2015) which express visual sentiment of video frames on a scale of 1 (negative) to 5 (positive). We sample frames at regular intervals and compute the affect evoked by these frames using this 5-point scale. Another higher level feature we consider is the presence of faces, which has previously been shown to have a strong influence on likes and comments in image-based social media (Bakhshi and others 2014). We therefore adapt it to the video context by computing the *fraction of frames with faces*. Finally *Number of past posts* by the creator of the video under consideration is also included to reflect user experience and activity on the social media network.

Social features We consider the *number of followers* of author of a content as a direct feature to reflect the user's social network capital. A more detailed list can be seen in Table 2

Micro-videos as a new form of expression

VISUAL QUALITY

	YT	Vine	Flickr
YT	0.00	3.31	5.17
Vine		0.00	5.39
Flickr			0.00

AUDIO

	YT	Vine	Flickr
YT	0	7.37	
Vine		0	
Flickr			

OBJECTS

	YT	Vine	Flickr
YT	0	1.18	2.26
Vine		0	1.51
Flickr			0

Figure 2: *Aggregated symmetric KL Divergences between platforms over all features for different feature groups.*

To understand the role of Vine and the different features of its videos, we begin by comparing Vine with image-based platform Flickr and video platform YouTube, and test the conjecture that Vine is a new form of expression that is different from both images and videos (Pogue 2013). We focus on items which have engaged large numbers of users, and compare our dataset of popular Vines (POP12K) with datasets of popular images from Flickr and videos on YouTube[9].

We compare the features (See Table 2) of POP12K vines with the corresponding features from a sample of 1000 popular[10] images from the MIR-Flickr dataset (Huiskes and Lew 2008), and 419 viral YouTube videos (Jiang and others 2014). To compare all the datasets as images, we sample the videos at the rate of 1 image per second producing several still images, whose parameters we study. We compare these datasets along three dimensions as below.

Our methodology is straightforward: For a given platform and feature, we can compute the marginal probability distribution of the feature across all videos (images) of the platform. For a given platform and feature, these distributions allow us to compute the probability of a random video (image) from that platform having a particular feature value. For a given feature, we can measure the differences between any pair of platforms by computing the symmetric Kullback-Liebler (K-L) divergence between the corresponding distributions of the feature. For a given pair of platforms, the average of the K-L divergence values across all features tells us how far apart the two platforms are in feature space. Fig. 2 shows the differences between the platform across the visual, audio and object (or visual semantic) features described previously (Table 2). We discuss this further by focusing on the features with the highest K-L divergence across each category of features:

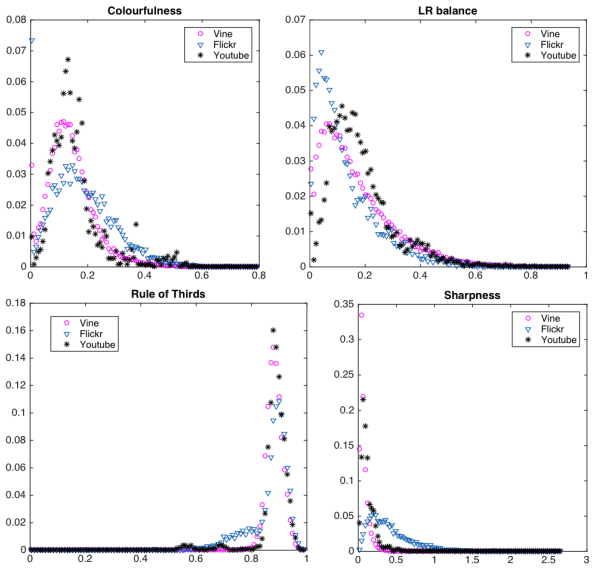


Figure 3: *Distributions of the most diverging aesthetic features across platforms.*

Visual quality comparison As expected, we find (Fig. 2) that Vine videos show different aesthetic behaviour from

[9]Similar results were obtained replacing the popular videos with videos from the ALL120K dataset, but are not shown here.

[10]i.e., with high "interestingness" rating – https://www.flickr.com/explore/interesting rating

Features	dim	Description
Visual Quality Features		
RMS contrast	1	RMS contrast is calculated as standard deviation across all the pixels relative to mean intensity
Weber Contrast	1	Weber contrast is calculated as $F_{weber} = \sum_{x=width} \sum_{y=height} \frac{I(x,y)-I_{average}}{I_{average}}$
Gray Contrastv	1	Gray contrast is calculated in similar to RMS contrast in HSL colour space for the L value of pixels.
Simplicity	2	Simplicty of composition of a photograph is a distinguishible factor that directly correlates with professionalism of the creator (Ke, Tang, and Jing 2006). We calculate Image simplicity by two methods: Yeh simplicity (Yeh 2010) and Luo simplicity (Luo and Tang 2008).
Naturalness	1	How much does the image colors and objects match the real human perception?To compute image naturaleness we convert the image into the HSV color space and then identify pixels corresponding to natural objects like skin, grass, sky, water etc. This is done by considering pixels which an average brigtness V $\in$ [20 , 80] and saturation S > 0.1. The final naturalness score is calculated by finding the weighted average of all the groups of pixels. (Zhong and others 2015).
Colourfulness	1	A measure of colourfulness that describes the deviation from a pure gray image. It is calculated in RGB colour space as $\sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{yb}^2}$ where $rg = R - G$ and $yb = \frac{R+G}{2}$ and μ and σ represent mean and standard deviation respectively
Hue Stats	2	Hue mean and variance which signifies the range of pure colours present in the image. It is directly derived from the HSL colour space
LR balance	1	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 amateur 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.
Rule of Thirds	1	This feature deals with compositional aspects of a photograph.This feature basically calculates if the object of interest is placed in one of the imaginary intersection of lines drawn at approximate one third of the horizontal and vertical postions. This is a well known aesthetic guideline for photographers.
ROI proportion	1	Measure of prominance given to salient objects. This measure detects the salient object in an image and then measures proportion of pixels its relative to the image
Image brightness	3	Features signify brightness of the image. Includes average brightness, saturation and saturation variance
Image Sharpness	1	A measure of the clarity and level of detail of an image. Sharpness can be determined as a function of its Laplacian normalized by the local average luminance in the surroundings of each pixel, i.e. $\sum_{x,y} \frac{L(x,y)}{\mu_{xy}}$, with $L(x,y) = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2}$ where $\mu_x y$ denotes the average luminance around pixel (x, y).
Sharp Pixel Proportion	1	Out of focus or blurry photographs are generally not considered aesthetically pleasing. In this feature we measure the proportion of sharp pixelscompared to total pixels. We compute sharp pixels by converting the image in the frequency domain and then looking at the pixel corresponding to the regions of highest frequency (Yeh 2010), using the OpenIMAJ (Hare and others. 2011) tool.
Higher Level Features		
Face Percentage	1	Percentage of frames in a video, which have been tested positive for atleast one face. Faces detected using Viola Jones Detector (Viola and Jones 2004)
Frame sentiment	1	Median frame sentiment of all the sampled frames from a micro video. The sentiment was calculater using the Multilingual Visual Sentiment Ontology detector (Jou and others 2015)
Past post count	1	Number of past posts user has uploaded prior to current one. This is a good measure of user's experience with the platform and activity.
Audio Features		
Zero Crossing rate	1	Zero crossing rate measures the rhythmic component an audio track (Laurier et al. 2009). It ends up detecting percussion instuments like Drums in the track
Loudness	2	This feature expresses overall percieved loudness as two components. Overall energy and average short time energy (Lartillot and Toiviainen 2007)
Mode	1	This feature estimates the musical mode of the audio tract (major or minor). In western music theory, major modes give a perception of happiness and minor modes of sadness. (Laurier et al. 2009)
Dissonance	1	Consonance and dissonance in an audio track has been shown to be relevent for emotional perception (Laurier et al. 2009). The Values of Dissnonance are a calculate by measuring space between peaks in the freaquency spectrum of the audio track. Consonant frequency peaks tend to be spaced evenly where as dissnonant frequency peaks are not
Onset Rate	1	This measures the the Rhythmical perception. Onsets are peaks in the amplitude envelop of a track. Onset rate is measured by counting such events in a second. This typically gives a sense of speed to the track.
Social Features		
Followers	1	Number of followers that the user posting a video has. This is the prime social feature available from the user metadata. The number of followers direcltly represent the audience which are highly probably to engage with the video on upload.

Table 2: Dimensionality and description of features used to describe Vine videos

both Flickr and YouTube, although they are more stylistically similar to YouTube videos than Flickr images. Taking a closer look at the features with the highest KL divergence values across all platform pairs, we notice that, in practice, Flickr aesthetic features reflect the behaviour of a "professional" medium. At the other end 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. More specifically, we notice the following (See Fig. 3). (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:* Some Flickr pictures tend to deviate from the standard rule of thirds, as it occasionally happens for professional, artistic pictures (Freeman 2007). On the other hand, the moving images of Vine and Youtube tend to stick to the Rule of Thirds heuristic. (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 professional expertise of Flickr photographers is clearly exposed by the higher percentage of sharp pixels.

Audio Channel Comparison. The audio channel is as important as the visual channel for long viral videos[11]. Fig. 2, we see that in the audio space Vine and YouTube are far apart, with the highest K-L divergence value across all categories of features and all platform pairs (Note that the Flickr image dataset does not have an audio component). 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, au-

[11] http://thenextweb.com/socialmedia/2015/03/20/set-the-tone-the-importance-of-sound-in-viral-videos/

dio 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. 4.

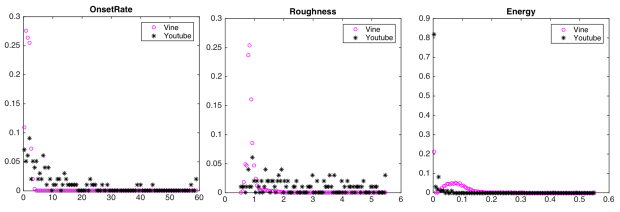


Figure 4: *Distributions of the most diverging audio features across platforms.*

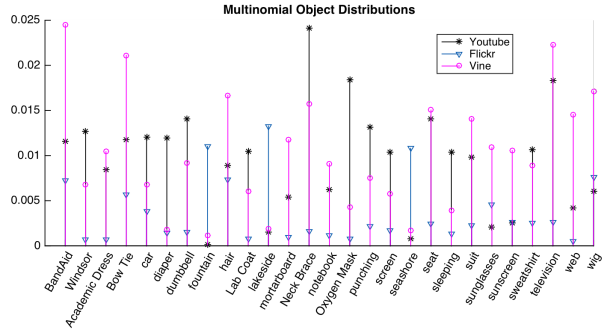


Figure 5: *Distributions of the most distant object occurrences across platforms, showing a preoccupation with fun and entertainment on Vine, 'traditional' popular subjects such as kids, cars and violence on YouTube, and photographic scenery on Flickr.*

Visual Semantics Comparison. Finally, we examine platform differences in terms of the visual objects they focus on. For each image in Flickr, and each sequence of images sampled from the YouTube and Vine videos, we use deep learning-based object detectors from ImageNet (Krizhevsky and others 2012), and 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. 2, 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 the top 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 and entertainment (*academic dress*, *wig*, *tv*, *sunglasses*). Viral Youtube videos prefer popular subjects such as kids (*diaper*), cars, and violent scenes (*punching*, *neck brace*). Finally, Flickr pictures can be distinguished with the presence of visual concepts typical of photographic sceneries (*lakeside*, *seashore*). These results could be taken to indicate that the three platforms are used for different purposes, with micro-videos being more of an immediate medium, 'capturing the moment' as it happens. This may help explain the nature of user engagement we observe in the rest of the paper.

User Engagement in Vine

The above preliminary analysis indicates that Vine has characteristics that distinguish it from popular image- and video-based platforms. Next we explore how these features affect user engagement by designing a novel but straightforward methodology to understand what features are important: We build a machine learning model that predicts engagement with high accuracy, precision and recall based on our audiovisual features, and use the relative feature importances to discuss the importance of features.

Metrics and methodology

To understand which aspects or features are important for user engagement, we need to: (*i*) define a metric for engagement, and (*ii*) develop a methodology to study how the metric is influenced by different features.

Defining a metric for user engagement: In this paper, we use *number of loops* of a micro video as a proxy for user engagement towards it[12]. Although user engagement is a broadly used term, and other metrics may well be used to represent user engagement, our choice is in line with previous related social media studies (e.g. (Bakhshi and others 2014)) that have used social attention metrics such as likes and comments to study user engagement. Video hosting platforms like Youtube also use the number of views (similar to number of loops on Vine) as a core metric for their user engagement API[13]. In the rest of the paper, we will use popularity and engagement interchangeably.

Motivating the methodology: Given a set of features, if we can build a machine learning model that uses the features to predict which content items are highly engaging, the relative importance of the different features in making the prediction can tell us about the relationship between the features and engagement. However, our results will only be as 'good' as the model is in predicting loop counts. Since predicting popularity with exact numbers such as loop counts is a hard problem, we turn to a simpler one: We define an arbitrary threshold count for loops, and bin micro videos as popular

[12]We obtained similar trends using number of reposts, but only report results with loops. Note that the loop counts of videos are highly correlated with reposts and likes. For example for videos in POP12K, $corr(Loops, Likes) = 0.80$, $corr(Likes, Reposts) = 0.91$, $corr(Reposts, Loops) = 0.74$.

[13]https://developers.google.com/youtube/analytics/v1/dimsmets/mets

or unpopular depending on whether the loop count is over or under the threshold.

We then design a classifier that predicts whether a micro videos is popular or unpopular (alternately, as engaging or not) based on our set of 28 features (Table 2) . As discussed next, a simple random forest classifier can be trained to make this prediction with high precision and accuracy. The relative importance of different features then tells us about how the features affects user engagement.

This method has one major limitation: its dependence on the arbitrarily defined loop count threshold. Therefore, we conduct a a sensitivity analysis by training a series of binary classifiers for different loop count thresholds. This also allows us to study shifts in relative importance, as we move up the scale towards more popular and engaging objects, by defining increasingly higher numbers of loop counts as the threshold for categorising a video as popular (or engaging).

Model details

Setup We sample 12,000 videos from our dataset, out of which 6,000 are popular videos from POP12K, and 6,000 randomly sampled from the ALL120K dataset, thus representing the entire spectrum of engagement levels. In each video, we sample the video track for individual frames at every second, and extract the audio track as well as metadata related to the video and its author. Using these, we then compute the 28 dimensional vector of all the features in Table 2 and train a random forest classifier to distinguish popular and unpopular videos for different thresholds of popularity. We used the implementation from the *SKLearn* package with $\sqrt{n_{features}}$ split and 500 estimators, which provided the best tradeoff between speed and prediction performance.
Performance Results Different classifiers are trained using the above method for different engagement/popularity thresholds, using an 80-20 split for training and testing. Fig 6c shows how these perform as we vary the threshold of "engagement" (popularity) from 80 loops (the median for ALL120K) to $\approx$ 500,000 loops (1.5 times the median of the popular videos i.e., POP12K). At each training iteration with a changed "engagement" threshold, we re-balance the dataset by choosing equal number of samples which fall in either classes. The classifiers gave consistently high performance (see lines labeled 6 sec), never dropping below 90% for accuracy, and 80% F-1 score, validating our next results about the importance of different features.

Feature analysis and implications

The impact of individual features on user engagement is calculated using Gini importance (Louppe and others 2013), and combined into social and content-related (i.e., audio and video-related) features as described before (§). Fig. 6a shows the trends in feature importance as a function of engagement threshold used (see lines labeled 6 sec). We observe that at lower thresholds of popularity, social features are much more important than content-related features, but at higher thresholds, content-related features increase in importance to become just as important as social features, suggesting that *content quality is important for user engagement at the top end of engagement*.

We drill down further in Fig. 6b, and examine the importance of different kinds of content-related features. For each class of content-related features, we plot the mean of the feature set of the class. We observe that in terms of effective importance of different feature tracks, sentiment is the weakest influencer in the classifier decision process. We conjecture that the relative lack of importance of sentiments may partly be due to the extremely short nature of micro videos, which does not let emotional 'story arcs' and plots (e.g., drama) to develop as strongly as in longer videos.

Further, we observe that the presence of faces in a frame strongly outweighs all other content-related features in predicting popularity. We confirm this in Figure 7 by comparing the percentage of faces in popular POP12K videos with the corresponding percentage in ALL120K videos (which contain a large number of unpopular videos as well as a few popular ones). These results indicate that popular videos tend to have more faces, i.e., "*faces engage us*". This is in alignment with similar results on other platforms, which also indicate that faces greatly enhance popularity related metrics such as likes and comments (Bakhshi and others 2014).

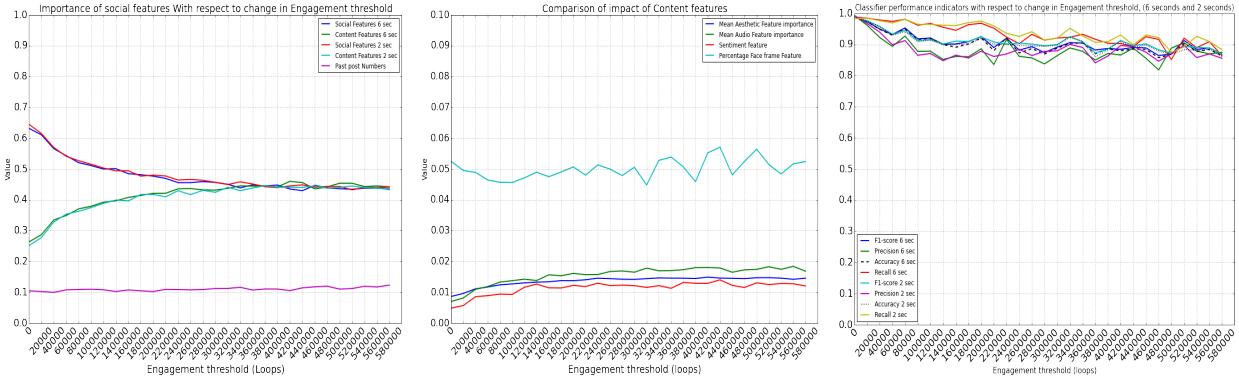
Primacy of the first seconds

Our main observations so far are: (*i*) Vine micro videos appear to be different from both images (Flickr) and videos (YouTube) and (*ii*) content quality features, especially image-related ones (e.g., fraction of faces with frames) become important for user engagement and popularity metrics. In this section, we try to understand these findings further: One way to think about videos is as a sequence of images. With micro videos, this sequence is of course much shorter than in other videos. To see whether the visual content of micro videos is more image-like or video-like, we consider the impact and evolution of image-related features from the beginning to the end of the micro video.

Image quality deteriorates over time

Vine videos can be at most 6.5 seconds long. We sample the videos twice every second and represent the whole video as a series of 12-13 static frames. This sampling rate is not too low to miss any considerable frame transitions, neither is it too high to include a lot of mid transition frames. For each sampled frame, we calculate the feature under consideration – sentiment, percentage of faces, and aesthetic score. To compute the aesthetic score, we extract the 18 aesthetic features described in Table 2. for each frame frame. To find an aggregate overall aesthetic score of each frame, we use a weighted sum of all the features (This is possible because all the features are on the same scale), where the weights are calculated to be proportional to the importance of each feature in the classifier designed in the previous section.

For each video and each feature, we then compute when in the video the feature reached its maximum value. We then divide the videos into two second intervals, essentially dividing the video into its first third, second third and third third. We then ask what proportion of videos had the maximum value of a feature in the first (respectively second and third) third. This procedure tells us when we are likely to find the 'best' part of the video.



(a) Content Vs Social Features (b) Individual Content Features (c) Classifier Performance

Figure 6: Understanding engagement for different thresholds (min. number of loops considered as engaging). Two different classifiers are used, one using quality of the entire micro video (labelled 6 sec), the second measuring quality from only the first two seconds (labelled 2 sec). (a) As threshold becomes higher, content-related factors become as important as social factors (both classifiers). Note that unlike content quality computed from the first 2 seconds ('Content features 2 sec') rather than the entire 6 seconds of the video ('Content features 6 sec'), 'social features 6 sec' uses the same information as Social Features 2 sec', but the two are plotted separately to show the relative importance of social features in the 6 second vs 2 second classifier. (b) Amongst content features alone, presence of faces in the video is the single most dominant feature, across all threshold levels (6 second classifier) (c) Both 2 sec and 6 sec classifiers perform similarly across all metrics such as Precision, Recall and F1-score. Performance is high across all engagement thresholds: all metrics are consistently over 0.8 or 0.9.

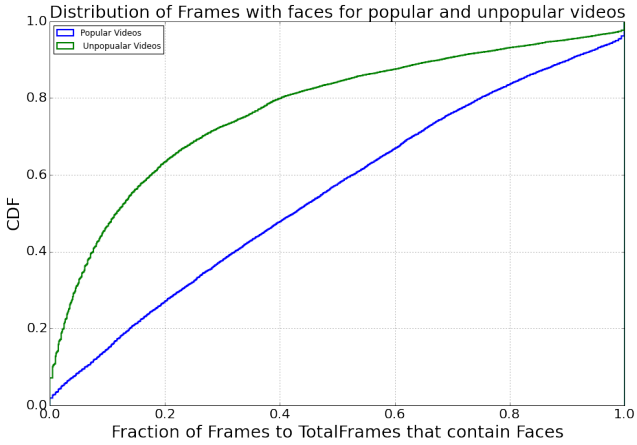


Figure 7: *CDF for popular and unpopular videos. The CDF signifies the cumulative distribution of percentages of frames containing faces in a vine video. The observation here is popular videos tend to have higher face percentage than unpopular videos*

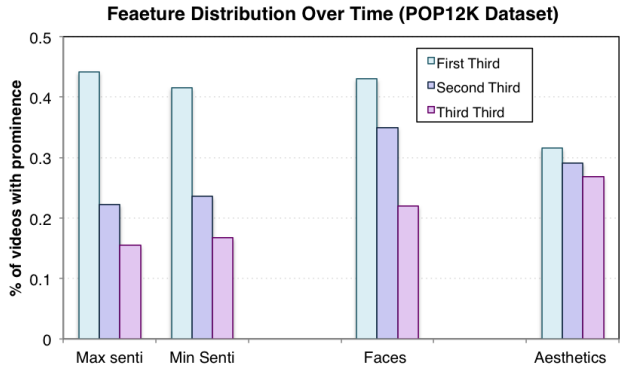


Figure 8: *Evolution of Feature magnitude: The graph shows sharp trend in prevalence of strongest component of a feature in the first one third of the video. The strength decreases progressively for the successive thirds. (Results shown for POP12K dataset. Similar results obtained for ALL120K.)*

Fig. 8 shows the result for each major category of content-related feature, plotted over both our datasets (ALL120K and POP12K). We observe a general trend where the first third has the maximum (best) value for all features considered. For instance, the best aesthetic score is to be found in the first two seconds. Similarly, the proportion of faces, an important predictor of engagement (Fig. 7), is also maximum in the first third.

Note that for sentiment values, the minimum value is just as valid and valuable as the maximum, representing a sad or emotionally dark segment of the movie with negative sentiment, in contrast to a happy segment of the movie with positive sentiment. Therefore, we calculate which third of the movie we find the maximum and minimum sentiment values and plot these separately. In both cases, we find yet again that the first third of the video has the maximum (minimum) sentiment value for the majority of videos.

Loops and likes are obtained on first sight: Initial seconds predict engagement

Collectively, the results above paint a picture where the first seconds of the micro video are highly important in engaging the user. We conjecture that this might be because of the mobile-first nature of Vine: the primary user interface is the Vine app, where users select which videos to watch by scrolling over it. The vine only plays when the user retains focus over the video, and hence the first seconds are likely critical for grabbing user attention and engaging them.

We next take this result to its logical conclusion, and ask how the classifier developed in the previous section for predicting engagement would work if using only content-related features from the first third of the video rather than from the whole micro video. Following the same methodology as described in the previous sub-section, we develop a series of classifiers for different popularity thresholds, training this time on image content-related features drawn from the first two seconds of the video rather than from across the whole video. The same set of hyper parameters were used as in the previous setting. As shown in Fig. 6c, the resulting classifiers (labeled 2 sec) perform very similarly to the classifiers developed before (labeled 6 sec). Further, Fig. 6a shows that the relative importance of different features is also nearly identical to the previous results. It should be emphasised that although these results were obtained using loop counts as the metric for user engagement, similar results have also been obtained using reposts (revines).

These results point to a *primacy of the first seconds* effect, whereby the first seconds of a micro video matter much more than the rest of the video, suggesting that they behave almost like still images in terms of user engagement.

Discussion and conclusions

In this paper, we took a first look at user engagement with micro videos. Defining engagement in terms of social attention metrics such as likes, revines (reposts) and loop counts, we find that content quality-related features have as strong an influence as social network-based exposure in driving these metrics. Furthermore, the quality of the first couple of seconds is higher than the quality of subsequent seconds, and can predict whether a micro video will be engaging or not, just as well as looking at the quality of the entire video. This 'primacy of the first seconds' effect makes the micro videos on Vine somewhat closer to user-generated images rather than user-generated videos, which we corroborated by comparing popular micro videos on Vine to popular images on Flickr and viral videos on YouTube.

This image-like quality of micro videos, and the importance of the video quality on popularity and user engagement has important implications and bearing on future work:

1. Advertisements on the Web are driven by social attention metrics. Therefore advertisers need to know and adjust their strategies based on the insight that that user attention is driven to a large extent by the first couple of seconds. Although in video ads do not appear to be common in today's micro videos, how to place ads that grab user attention within a short duration of time will be a problem that is interesting both from a research as well as business perspective.
2. A possible reason for the deterioration of image quality is that it may be difficult to maintain image composition, focus etc using a mobile phone camera with moving subjects. Novel UI and multimedia techniques that can help correct for such quality deterioration could greatly help micro video creators and is a second fruitful direction for further study.
3. Recently, several micro video platforms have started extending the duration of micro videos. Although the wisdom of longer micro videos without appropriate editing tools has been questioned[14], from a research perspective it would be interesting to study how user behaviour and engagement changes as longer micro videos become more common place. Interestingly, we find that in a small sample of about 4,500 Instagram videos (where the maximum permitted duration is 60 seconds), users continue to prefer shorter videos, with 70% of videos less than 20 seconds long, and the median duration at just under 15 seconds.

More generally, in this work we considered user engagement as a single dimension. However, we acknowledge that user engagement is a very subjective notion, impacted by different factors including user location, habits, gender, visual preferences. In the future, we plan to explore how such different user sub-cultures perceive and engage with micro videos, following recent works from the Multimedia community studying the impact of culture in subjective image perception (Jou and others 2015). A second dimension to explore in our future work is generalising the above findings to other micro video platforms – our preliminary studies indicate that key results such as the Primacy of the first seconds effect, are robust across platforms, applying to Instagram videos as well. However, more work is required in this direction.

References

Bakhshi, S., et al. 2014. Faces engage us: Photos with faces attract more likes and comments on instagram. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, 965–974. ACM.

Cha, M., et al. 2009. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th WWW*, WWW '09, 721–730. New York, NY, USA: ACM.

Chen, J., et al. 2016. Micro tells macro: Predicting the popularity of micro-videos via a transductive model. In *Proceedings of the 2016 ACM on Multimedia Conference*, MM '16, 898–907. New York, NY, USA: ACM.

Datta, R., et al. 2008. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *2008 15th IEEE ICIP*, 105–108. IEEE.

et.al, K. 2015. How to take a good selfie? In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, 923–926. New York, NY, USA: ACM.

[14] http://www.theverge.com/2013/6/20/4448906/video-on-instagram-hands-on-photos-and-video

Fontanini, G., et al. 2016. Web video popularity prediction using sentiment and content visual features. In *Proceedings of the 2016 ACM on ICMR*, 289–292. ACM.

Freeman, M. 2007. *The Photographer's Eye: Composition and Design for Better Digital Photos*, volume 1. Focal Press.

Grossman, D. 2013. Can micro video change how we communicate? BBC Newsnight.

Hare, J. S., and others. 2011. Openimaj and imageterrier: Java libraries and tools for scalable multimedia analysis and indexing of images. In *Proceedings of the 19th ACM MM*, MM '11, 691–694. New York, NY, USA: ACM.

Huiskes, M. J., and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM ICMR*. New York, NY, USA: ACM.

Isola, P.; Xiao, J.; Torralba, A.; and Oliva, A. 2011. What makes an image memorable? In *CVPR, 2011 IEEE Conference on*, 145–152. IEEE.

Jiang, L., et al. 2014. Viral video style: A closer look at viral videos on youtube. ICMR '14. New York, NY, USA: ACM.

Jou, B., et al. 2015. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM MM*, 159–168. ACM.

Ke, Y.; Tang, X.; and Jing, F. 2006. The design of high-level features for photo quality assessment. In *2006 IEEE CVPR'06*, volume 1, 419–426. IEEE.

Khosla, A., et al. 2014. What makes an image popular? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14.

Krizhevsky, A., et al. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, 1097–1105.

Lartillot, O., and Toiviainen, P. 2007. A matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*, 237–244.

Laurier, C.; Lartillot, O.; Eerola, T.; and Toiviainen, P. 2009. Exploring relationships between audio features and emotion in music.

Leung, L. 2009. User-generated content on the internet: an examination of gratifications, civic engagement and psychological empowerment. *New media & society* 11(8):1327–1347.

Louppe, G., et al. 2013. Understanding variable importances in forests of randomized trees. In *NIPS*, 431–439.

Luo, Y., and Tang, X. 2008. Photo and video quality evaluation: Focusing on the subject. In *European Conference on Computer Vision*, 386–399. Springer.

Machajdik, J., and Hanbury, A. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM MM*, MM '10. New York, NY, USA: ACM.

Mazloom, et al. 2016. Multimodal popularity prediction of brand-related social media posts. In *Proceedings of the 2016 ACM on Multimedia Conference*, MM '16.

Nguyen, P. X.; Rogez, G.; Fowlkes, C.; and Ramamnan, D. 2016. The open world of micro-videos. *arXiv preprint arXiv:1603.09439*.

Nwana, A. O., et al. 2013. A latent social approach to youtube popularity prediction. In *2013 IEEE (GLOBECOM)*, 3138–3144. IEEE.

O'Brien, H. L., and Toms, E. G. 2008. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology* 59(6):938–955.

Pinto, H., et al. 2013. Using early view patterns to predict the popularity of youtube videos. In *Proceedings of the sixth ACM ICWSM*, 365–374. ACM.

Pogue, D. 2013. Why are micro movies so popular these days? *Scientific American*.

Redi, M., and Others. 2014. 6 seconds of sound and vision: Creativity in micro-videos. In *Proceedings of the IEEE CVPR*, 4272–4279.

Schifanella, R., et al. 2015. An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures. In *Proceedings of THE 9TH ICWSM 2015*.

Schifanella, R., et al. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 2016 ACM MM*, 1136–1145. ACM.

Totti, et al. 2014. The impact of visual attributes on online image diffusion. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14. New York, NY, USA: ACM.

Viola, P., and Jones, M. J. 2004. Robust real-time face detection. *International journal of computer vision* 57(2):137–154.

Wang, Y., et al. 2015. Inferring sentiment from web images with joint inference on visual and social cues: A regulated matrix factorization approach. In *Ninth ICWSM*.

Yamasaki, T., et al. 2014. 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*, WISMM '14. New York, NY, USA: ACM.

Yeh, C.-H. e. a. 2010. Personalized photograph ranking and selection system. In *Proceedings of the 18th ACM MM*, 211–220. ACM.

Zhong, C., et al. 2015. Predicting pinterest: Automating a distributed human computation. In *Proceedings of the 24th WWW*, WWW '15. New York, NY, USA: ACM.