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

Sagar Joglekar[1], Nishanth Sastry[1], Miriam Redi[2],

[1] Kings College, London, UK
[2] Bell Labs, Cambridge, UK

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 ($> 120,000$) 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 verify these data-driven insights with a user study from 115 respondents, confirming that users tend to engage with micro videos based on "first sight", and that users see this format as a more immediate and less professional medium than traditional user-generated video (e.g., YouTube) or user-generated images (e.g., Flickr).

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 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[3]), the extremely short duration time limits of Vine etc has led to the coining of a new term: *micro videos*. Some media commentators have argued that the restrictions imposed by the micro video format could fundamentally change the way we communicate [7]. Indeed, it has been argued that Vine has had a significant

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

cultural impact far beyond its user base, generating several widely shared memes in its short lifetime[4].

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[5]. 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[6].

This paper aims to examine how crucial changes such as social network integration and time limit expansion might affect user engagement with this new format. To better understand these issues, we formulate the following research questions:

RQ1 What are the relative roles of social and content quality factors in driving engagement and popularity in micro-videos?

RQ2 How does the strict time limit impact video quality, and user engagement (both as creators and consumers) with such videos?

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 micro-video platforms such as Instagram[7].

To address **RQ1**, 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 [1], we find that presence of faces in vine videos significantly increases engagement, and is the most important content-related factor.

Next, to explore **RQ2** we look at how the quality of the video varies 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 (sentiment) space, are

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

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

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

[7] On acceptance, our newly crawled data will also be shared for non commercial research

more prevalent in the initial seconds of the micro video, suggesting that the authors are consciously or subconsciously 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.

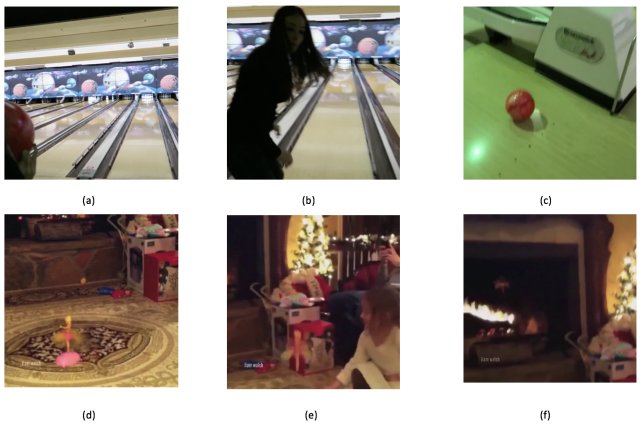


Fig. 1: *Vine Samples from first, second and thirds one thirds of the video. Images (a) , (b) and (c) show a progressive drop in brightness and sharpness due to shaky camera. Images (d) (e) and (f) shows a progressive drop in contrast.*

We confirm these computationally acquired findings with real user impressions by designing a survey which was answered by 115 respondents: Over 66% of users react to (like/comment on) content from their friends, making social interaction a significant part of content consumption; and 44% of users form opinions about videos in the first few seconds, validating the observed primacy of first seconds effect. Our survey also suggests that platforms such as Vine are seen as less professional and more immediate formats than, say Flickr images or YouTube videos, providing support to David Pogue's position that micro videos are a new kind of user-generated content [25], and therefore should be treated differently when it comes to user engagement.

2 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 [13], researchers have started

exploring concepts beyond semantics, such as image memorability [9], emotions [19], and, more broadly, pictorial aesthetics [4, 18, 5]. 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 [26], sentiment [32, 10] and sarcasm [28].

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 [35]; recent work [20] 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, [33] used textual annotations to predict various popularity metrics of social photos. Social metrics such as early views [24] or latent social factors [22] have also been used to effectively estimate video popularity. However, the fact that many popular media items may not depend on the social network [2] 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: [30] analysed the visual attributes impacting image diffusion, and [27] studied relations between image quality and popularity in on-line photo sharing platforms. Bakhshi et al [1] 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 [12]. Moreover, a work done by Fontanini et.al [6] explores the 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.

Micro videos are relatively new, so work specifically on micro video analysis has been limited. Redi *et al.* [26] quantify and build on the notion of creativity in micro-videos. A large dataset of 200K Vine videos was collected by Nguyen et.al [21], focusing on analysis of tags. Closest to our work is Chen *et al.* [3] 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.

3 Introduction to Datasets

Micro videos were pioneered and popularised by Vine[8], which was launched in 2012. Vine videos are constrained to a maximum length of 6.5 seconds. Videos are typically created using the mobile app and posted on user's profile which can be followed and shared by other users within the app or the website. We stress most of our work on videos sampled from Vine, complemented by Instagram data, which will be introduced as appropriate. The rest of this section gives details about the Vine datasets.

3.1 Dataset description

Dataset	Posts (total)	Loops/Views (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[9] 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.

[8] http://vine.co

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

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 after initial 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.

3.2 Feature Descriptions

In order to fully understand how micro-videos engage users, we characterize the content of videos using computer vision and computational aesthetics techniques and extract a number of features (Table 3 in Appendix), which can be divided into the following categories:

Image quality features These features are mostly taken from computational aesthetics literature, and have been recognized as heuristics for good photography. Prior work [35] 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 with higher 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 [26], we use audio features known to have an impact on emotion and reception. Using open source tools [14, 15], we measure *loudness* (overall volume of the sound track), the *mode* (major or minor key), *roughness* (dissonance in the sound track), and *rhythmical* features describing abrupt rhythmical changes in the audio signal.

Higher Level features Affect (emotions experienced) is well known to strongly impact on user engagement [23, 16]. To understand the sentiment conveyed by the video frames, we use the Multi Lingual Sentiment Ontology detectors [10] 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 [1]. 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 the author of a content as a direct feature to reflect the user's social network capital.

A more detailed description of all the features can be seen in Table 3.

4 User Engagement in micro videos

We begin our analysis by devising a novel methodology to analyze how the previously defined features impact user engagement in micro videos (**RQ1**). Our results indicate the importance of social features for highly engaging videos, and that the presence of faces is a strong content-related feature that positively impacts user engagement.

4.1 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[10]. 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. [1]) 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[11]. 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 categorize micro videos as popular 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 3) . As discussed next, a simple random forest classifier can be trained to make this prediction with high

[10] 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, $corr(Loops, Likes) = 0.80$, $corr(Likes, Reposts) = 0.91$, $corr(Reposts, Loops) = 0.74$.

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

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 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 categorizing a video as popular (or engaging).

4.2 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 3 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 trade-off 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 validation. Fig 2c 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. We take care that we are training on at-least 20% of the complete dataset by the end of the process, and stop increasing the threshold beyond that point to avoid over-fitting. The classifiers gave consistently high performance on the validation dataset (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.

4.3 Feature analysis and implications

The impact of individual features on user engagement is calculated using Gini importance [17], and combined into social- and content-related (i.e., audio and video-related) features as described before (§3.2). Fig. 2a 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*. This facet of users' engagement with Vine might legitimize Twitter's decision to more closely integrate the Vine platform with its social network: since a large part of micro-video popularity can be

explained with social factors, a better social network might further foster engagement with this unique form of expression.

We drill down further in Fig. 2b, 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 3 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 [1].

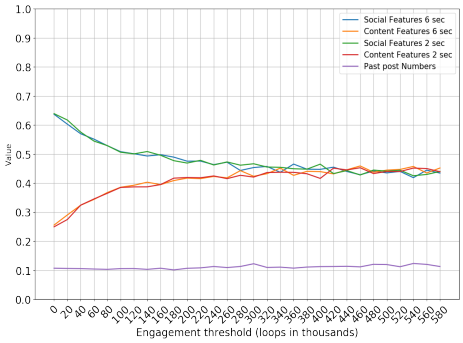
5 Primacy of the first seconds

Next, we try to understand these findings further by examining the quality of the individual frames of the videos: 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, and we investigate whether this has impact on video quality (**RQ2**). Our results show a "primacy of the first seconds" effect, with quality deteriorating over time and the quality at the beginning is as good a predictor of engagement as quality of the entire video.

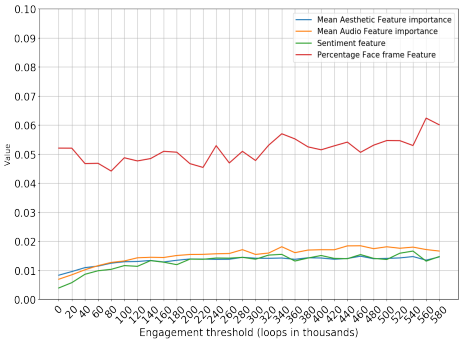
5.1 Image quality deteriorates over time

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 3. 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.

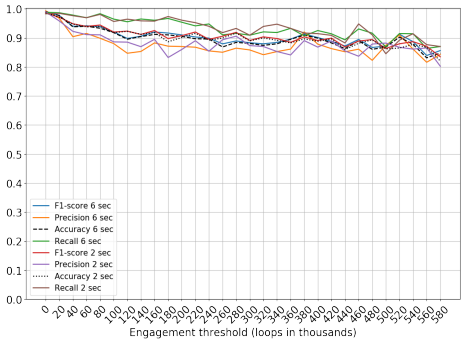
For each video and each feature, we then compute when in the video the feature reached its maximum value. We then divide the videos into two second intervals, essentially dividing the video into its first third, second third and third third. We then ask what proportion of videos had the maximum value 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

Fig. 2: 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 (labeled 6 sec), the second measuring quality from only the first two seconds (labeled 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 feature values 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.

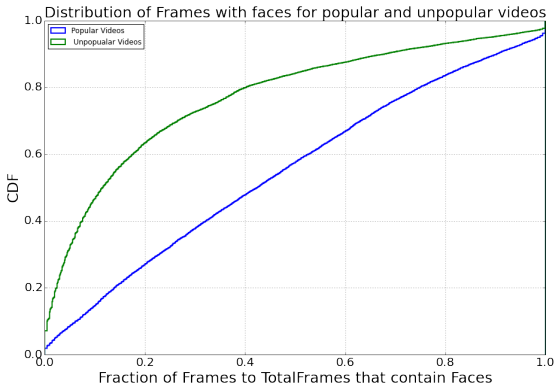


Fig. 3: *CDF for popular and unpopular videos. The CDF signifies the cumulative distribution of percentages of frames containing faces in a vine video. The observation here is popular videos tend to have higher face percentage than unpopular videos*

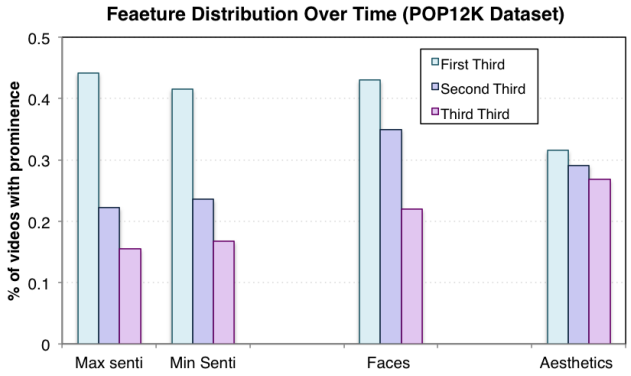


Fig. 4: *Evolution of Feature magnitude: The graph shows sharp trend in prevalence of strongest component of a feature in the first one third of the video. The strength decreases progressively for the successive thirds. (Results shown for POP12K dataset. Similar results obtained for ALL120K.)*

Fig. 4 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. 3), 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.

5.2 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. 2c, the resulting classifiers (labeled 2 sec) perform very similarly to the classifiers developed before (labeled 6 sec). Further, Fig. 2a shows that the relative importance of different features is also nearly identical to the previous results. It should be emphasized that although these results were obtained using loop counts as the metric for user engagement, similar results have also been obtained using reposts (re-vines).

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

6 User study

To complement the data analysis and gain deeper insight into what drives user actions and engagement, we designed an anonymous user study which captures user behavior when engaging with micro-videos.

6.1 Survey methodology

We initially recruited undergraduate students, obtaining about 33 responses. Subsequently, we tweeted the survey out to the Official Vine Twitter account and to the accounts of Vine and Instagram users, in order to gain further exposure amongst users of these platforms. In all, 115 users responded to our survey. Table 2 summarizes the respondents' demography and usage preferences. Most questions asked were to be answered either on a 5-point Likert scale (Strongly agree to Strongly Disagree), and or in a semantic differential format, with three options to choose from.

Attribute	Value
Male respondents (%)	44.2
Female respondents (%)	55.8
Age Demography (%)	
18-24	43.4
25-31	34.5
32-40	14.2
40+	8

Table 2: Summary of survey responses

6.2 Validation of data-driven results

To understand engagement with micro-videos and validate the findings that emerged from our analysis of **RQ1** and **RQ2**, the survey asked the following 5 questions to be answered on a 5 level Likert scale (strongly agree - strongly disagree):

A I tend to like/comment on videos from friends rather than from strangers
B I always form an opinion of a video in the initial few seconds, once the video starts playing
C I rarely watch short videos (Snapchats, stories) , all the way to the end.
D I prefer to watch short videos of humans on these platforms. E.g. I like to see a person talking/expressing rather than outdoor scenery, or Cats.

Almost 66% users agreed to question **A**, which reaffirms the tendency of socially embedded users being able to get high engagement scores. 44% of users agreed to question **B** (and further $\approx$ 30% users were neutral) and 38% users agreed with statement **C**, supporting the observed the "primacy of the first seconds" effect. Contrary to our findings regarding faces, only 34% users agreed with statement **D** (although a further 39% remained neutral; thus only a minority 27% of users disagreed or disagreed strongly). Such result might suggest that for many users, our attraction towards face shapes is innate [29] and people do not consciously engage more with faces.

6.3 Understanding what matters to users

The next part of the survey went beyond confirming the data-driven analysis by asking the respondents how their behavior changed when it comes to *acting* on a video they engaged with, i.e., when do they like/forward, comment or stop playing the video? 44% like, comment or share videos only after finishing watching it, and but a sizable 56% agreed that they do so in the middle of watching the video itself, or right at the beginning (19% share at the beginning. 37% somewhere in the middle); again pointing to the need for capturing users in the initial parts of the video.

Interestingly, a majority of 55% of respondents agreed (on a 5 point Likert scale) to the statement: "I don't really care about the quality of the micro-video or stories, as long as I like the content". This result, together with the previous answers seems to imply that the fall off in content quality in the latter parts of micro videos does not negatively impact user engagement. However, users do see a difference between micro videos and "traditional" (and older) user generated content platforms such as YouTube:

an overwhelming 75% of respondents rate the production quality of YouTube videos quality as more professional than micro videos.

7 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. We further conduct a user study to understand ground-truth user behavior when it comes to micro-videos. The study suggests that users tend to make quick opinions regarding micro-videos and engage with them almost in an image-like fashion, where they may begin but not finish watching the short 5-10 second long video.

These aspects of micro video 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 user attention is driven to a large extent by the initial seconds. Although 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 and a 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 also represent a second promising 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[12], from a research perspective it would be interesting to study how user behavior and engagement changes as longer micro videos become more common place. Interestingly, we find that in a small sample of about 6000 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. Such user preferences can and should be considered as the micro video format evolves further on different platforms.

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 future work, we plan to explore how such different user sub-cultures perceive and engage with

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

micro videos, following recent works from the Multimedia community studying the impact of culture in subjective image perception [10]. 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.

A Appendix: Feature Table

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 Contrast	1	Gray contrast is calculated in similar to RMS contrast in HSL colour space for the L value of pixels.
Simplicity	2	Simplicity of composition of a photograph is a distinguishable factor that directly correlates with professionalism of the creator [11]. We calculate Image simplicity by two methods: Yeh simplicity [34] and Luo simplicity [18].
Naturalness	1	How much does the image colors and objects match the real human perception?To compute image naturalness 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 brightness 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. [35].
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 positions. This is a well known aesthetic guideline for photographers.
ROI proportion	1	Measure of prominence 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 pixels compared 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 [34], using the OpenIMAJ [8] tool.
Higher Level Features		
Face Percentage	1	Percentage of frames in a video, which have been tested positive for at-least one face. Faces detected using Viola Jones Detector [31]
Frame sentiment	1	Median frame sentiment of all the sampled frames from a micro video. The sentiment was calculated using the Multilingual Visual Sentiment Ontology detector [10]
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 [15]. It ends up detecting percussion instruments like Drums in the track
Loudness	2	This feature expresses overall perceived loudness as two components. Overall energy and average short time energy [14]
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. [15]
Dissonance	1	Consonance and dissonance in an audio track has been shown to be relevant for emotional perception [15]. The values of dissonance are a calculate by measuring space between peaks in the frequency spectrum of the audio track. Consonant frequency peaks tend to be spaced evenly where as dissonant 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 meta-data. The number of followers directly represent the audience which are highly probably to engage with the video on upload.

Table 3: Dimensionality and description of features used to describe Vine videos

References

1. Bakhshi, S., et al.: 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. pp. 965–974. ACM (2014)
2. Cha, M., et al.: A measurement-driven analysis of information propagation in the flickr social network. In: Proceedings of the 18th WWW. pp. 721–730. WWW '09, ACM, New York, NY, USA (2009)
3. Chen, J., et al.: Micro tells macro: Predicting the popularity of micro-videos via a transductive model. In: Proceedings of the 2016 ACM on Multimedia Conference. pp. 898–907. MM '16, ACM, New York, NY, USA (2016), http://doi.acm.org/10.1145/2964284.2964314
4. Datta, R., et al.: Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In: 2008 15th IEEE ICIP. pp. 105–108. IEEE (2008)
5. et.al, K.: How to take a good selfie? In: Proceedings of the 23rd ACM International Conference on Multimedia. pp. 923–926. MM '15, ACM, New York, NY, USA (2015), http://doi.acm.org/10.1145/2733373.2806365
6. Fontanini, G., et al.: Web video popularity prediction using sentiment and content visual features. In: Proceedings of the 2016 ACM on ICMR. pp. 289–292. ACM (2016)
7. Grossman, D.: Can micro video change how we communicate? BBC Newsnight (Sep 2013)
8. Hare, J.S., others.: Openimaj and imageterrier: Java libraries and tools for scalable multimedia analysis and indexing of images. In: Proceedings of the 19th ACM MM. pp. 691–694. MM '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2072298.2072421
9. Isola, P., Xiao, J., Torralba, A., Oliva, A.: What makes an image memorable? In: CVPR, 2011 IEEE Conference on. pp. 145–152. IEEE (2011)
10. Jou, B., et al.: Visual affect around the world: A large-scale multilingual visual sentiment ontology. In: Proceedings of the 23rd ACM MM. pp. 159–168. ACM (2015)
11. Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: 2006 IEEE CVPR'06. vol. 1, pp. 419–426. IEEE (2006)
12. Khosla, A., et al.: What makes an image popular? In: Proceedings of the 23rd International Conference on World Wide Web. WWW '14, New York, NY, USA (2014), http://doi.acm.org/10.1145/2566486.2567996
13. Krizhevsky, A., et al.: Imagenet classification with deep convolutional neural networks. In: NIPS. pp. 1097–1105 (2012)
14. Lartillot, O., Toiviainen, P.: A matlab toolbox for musical feature extraction from audio. In: International Conference on Digital Audio Effects. pp. 237–244 (2007)
15. Laurier, C., Lartillot, O., Eerola, T., Toiviainen, P.: Exploring relationships between audio features and emotion in music (2009)
16. Leung, L.: User-generated content on the internet: an examination of gratifications, civic engagement and psychological empowerment. New media & society 11(8), 1327–1347 (2009)
17. Louppe, G., et al.: Understanding variable importances in forests of randomized trees. In: NIPS. pp. 431–439 (2013)
18. Luo, Y., Tang, X.: Photo and video quality evaluation: Focusing on the subject. In: European Conference on Computer Vision. pp. 386–399. Springer (2008)
19. Machajdik, J., Hanbury, A.: Affective image classification using features inspired by psychology and art theory. In: Proceedings of the 18th ACM MM. MM '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1873951.1873965
20. Mazloom, et al.: Multimodal popularity prediction of brand-related social media posts. In: Proceedings of the 2016 ACM on Multimedia Conference. MM '16, New York, NY, USA (2016), http://doi.acm.org/10.1145/2964284.2967210

21. Nguyen, P.X., Rogez, G., Fowlkes, C., Ramamnan, D.: The open world of micro-videos. arXiv preprint arXiv:1603.09439 (2016)
22. Nwana, A.O., et al.: A latent social approach to youtube popularity prediction. In: 2013 IEEE (GLOBECOM). pp. 3138–3144. IEEE (2013)
23. O'Brien, H.L., Toms, E.G.: 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 (2008)
24. Pinto, H., et al.: Using early view patterns to predict the popularity of youtube videos. In: Proceedings of the sixth ACM ICWSM. pp. 365–374. ACM (2013)
25. Pogue, D.: Why are micro movies so popular these days? Scientific American (May 2013)
26. Redi, M., Others: 6 seconds of sound and vision: Creativity in micro-videos. In: Proceedings of the IEEE CVPR. pp. 4272–4279 (2014)
27. Schifanella, R., et al.: An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures. In: Proceedings of THE 9TH ICWSM 2015 (2015)
28. Schifanella, R., et al.: Detecting sarcasm in multimodal social platforms. In: Proceedings of the 2016 ACM MM. pp. 1136–1145. ACM (2016)
29. Slater, A., Kirby, R.: Innate and learned perceptual abilities in the newborn infant. Experimental Brain Research 123(1-2), 90–94 (1998)
30. Totti, et al.: The impact of visual attributes on online image diffusion. In: Proceedings of the 2014 ACM Conference on Web Science. WebSci '14, ACM, New York, NY, USA (2014)
31. Viola, P., Jones, M.J.: Robust real-time face detection. International journal of computer vision 57(2), 137–154 (2004)
32. Wang, Y., et al.: Inferring sentiment from web images with joint inference on visual and social cues: A regulated matrix factorization approach. In: Ninth ICWSM (2015)
33. Yamasaki, T., et al.: 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, ACM, New York, NY, USA (2014)
34. Yeh, C.H.e.a.: Personalized photograph ranking and selection system. In: Proceedings of the 18th ACM MM. pp. 211–220. ACM (2010)
35. Zhong, C., et al.: Predicting pinterest: Automating a distributed human computation. In: Proceedings of the 24th WWW. WWW '15, ACM, New York, NY, USA (2015)