Study of Human affects in Vine videos

sagar , nishanth
King's College London, UK
sagar.joglekar, nishanth.sastry}@kcl.ac.uk

1. ABSTRACT

In the past few years we witnessed the rise of the *Selfie* phenomena. The world of online social networks was taken by a storm. There were several social, psychological and social computing studies trying to understand this phenomena. Our work defines and validates a new phenomenon of Video-selfie which behaves much more like a still selfie but in the video realm. The paper tries to measure video-selfies and understand the Face to frame ratios. The ask certain questions about the ontological description of video selfies and correlation of those with video popularity. Finally the paper proposes a new framework to understand video selfies using *Machine Intelligence* and validate it by benchmarking the predictors that arise from it. We do all our experiments on a crawled dataset gathered from the popular video service Vine, and then we do comparative analysis of a popular Instagram still selfie dataset.

2. INTRODUCTION

Online social networks (OSNs) have seen a massive surge in usage over the past decade. The surge is going hand in hand with the explosion of smart phone industry. More and more social interactions are now driven by media contents like selfies, group selfies and videos because of the ubiquitous nature of cameras. A sharp change in cultural aspects of online social interactions are evident and have also been studied in detail in papers like [12]. With the rise of social media networks like Vine and Instagram, human to human non-verbal interactions have another dimension to manifest. One of the predominant modality of self expression arose from this boom in social media, and that was the Selfie. The [12] paper explores several of the properties of selfie amongst Instagram users, where they explore correlation of facial orientation, poses and smiles with parameters like country of origin of the selfie user, post frequency, likes received , number of faces in the pictures, gender and smile scores. Such studies give us interesting insights about the sharp rising OSN phenomena of selfies. The study also states that more than 50 percent of photos shared on Instagram, fall under the category if selfies.

A major change in these behaviours was seen when the social network called Vine was launched in 2012. Vine adds another dimension to the act of self expression, where the users can record a 7 second long video and post it online and get engagement from peers. This service got so popular that the larger services like Instagram and Twitter also added the feature of short videos to their services. These developments just indicate that the fact that videos add more dimensions to the act of self expressions and they are being used heavily by the millennials and the young adult generation. This new modality of expression is least studied amongst academia and hence our paper tries to explore it and build a framework to understand it. We ask the following questions regarding this medium of expressions This looks good

1. Video Selfie: Is there something analogous to a still selfie in this new medium? How can we define it?
2. The topics: Can we do an ontological analysis of the types of videos that qualify as video-selfies?
3. Face matters: Is popularity of video selfie really correlated with the human faces?
4. Does it have a pattern: Can we extract any pattern information using new Machine learning techniques and predict the popularity numbers?

To our best of knowledge, this would be a first of its kind investigation into this type of medium of expression. We use Machine learning and Ontological extraction to build our framework. We work on real data collected from a popular social media service called Vine. We would make this dataset public post the study. In the following sections we will try to describe our problem and the solution approach.

3. RELATED WORK

There have been several attempts to understand the phenomena of self expression. With the rise of selfie, the expression exposed several facets of human nature. [12] looked at the phenomenon of selfie as a whole. There the authors explain social structures , temporal dynamics, demographics and memes using Instagram datasets of selfies. There were other works in this area [9] which explore the content itself. They talk about what kind of content is posted by specific categories of users. Also they try to understand how different types if content relate to the number and types of followers an account gathers. Another interesting work [10] goes in a different direction and tries to understand whys and whats of a perfect selfie. They try to analyse the ontological aspects of Instagram selfies using Sentibank [4] and discuss the salient characteristics of popular selfies. They also employ Deeplearning model of Imagenet [11] which can detect 1000 classes of objects in an image. These approaches help in understanding what are the most common themes of still selfies on Instagram. These works ask a bigger ques-

WWW 2015, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741671.

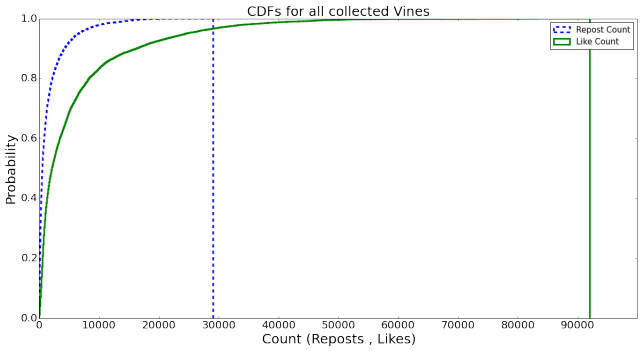

Figure 1: **CDF of likes and repost count across collected Vines.**

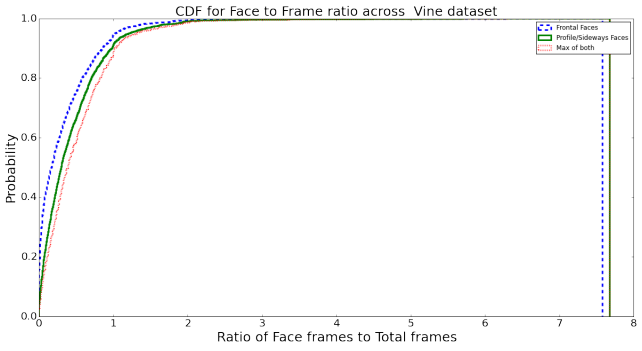

Figure 2: **CDF of Face frames to total frames calculated for Frontal and Profile faces**

tion in the realm of social media analysis. What aspects of media appeal to humans. What makes a media more or less consumable.

There were social media analysis of media in social network drawing conclusion on certain types of media gaining more popularity and engagement. The work by Bakshi et.al [1] shows that a study of instagram conclusively shows that photos with faces attract more likes and comments. We try to explore this hypothesis in videos. But first we would like to introduce some of our dataset numbers

4. DATASET

For our work we collected post metadata and videos URLs from Vine social network. We crawl for the top 100 popular posts over all and also over specific channel categories, every time we run the crawl cycle. We run the cycle about every 6 hours for over 1 month. We collect the metadata and to make sure there is minimal overlap in rankings and we continued this exercise for over a month. Finally we filtered all the uniques posts out and collected the actual vine mp4 files. In total we have 16504 unique vine clips collected over a month that ranked in the top 100 posts across vine. We also store the individual post metadata and the profile of the post creator. Some statistics for our dataset are as follows

Parameter	mean Value	Median Value
Reposts	1558	552
Likes	5754	2193
Loops	205504	76895

Figure 1 shows that the behaviour of reposts could be used as a viable metric for user's interaction with a vine video and dissemination of the video in the network. From the statistics of reactions to vines, it seems loops or the number of times a video is replayed is not necessarily a good measure to quantify popularity. However the likes and repost count could act as a good descriptor for a Vine's popularity. Finally we also reuse the selfie dataset collected by Kalayeh et.al [10], where they open 46000 selfie Instagram images to the public. We use this dataset to compare still selfie popularity behaviour with video selfie behaviour.

5. VIDEO SELFIE

The first question we ask about this dataset, to understand the contribution of human affects was, is there any such thing as a video selfie. The paper by Bakshi et.al [1] shows that faces in an image play a major role in popularity of that image on instagram. Our aim is to see whether a similar behaviour exist in video social networks like Vine. For this , we had to first plot some statistics regarding how many of the collected vines contain faces amongst the

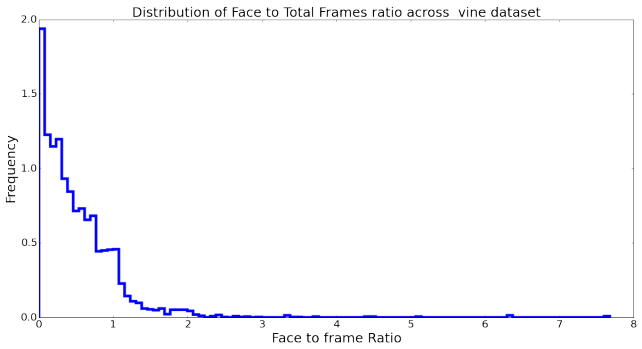

Figure 3: **Distribution of face frames compared to total frames for all the Vine videos in the dataset**

major fraction of the video frames. We would define these videos as video-selfies, or videos that potentially mimic still selfie like behaviour. To understand this better, we processed each video using the well known framework for face detection by Viola and Jones [13] to detect profile and frontal faces in a video. We collected these measurements across all the collected vine videos and stored the face frames as well as the statistics of each video. After plotting the CDFs of the ratio of face frames to total frames for all videos 2, we found that more than 50 percent of popular vines which ranked in top 100, had the ratio to be more than 0.5. The mean ratio of face frames to total frames was 54.74 percent, across all the collected samples. Further we plotted the distribution of these ratios as shown in 3. These numbers indicate that there is a similar phenomenon like still selfies , prevalent on video social networks like vine.

6. FRAME CONTENT ANALYSIS

With the hypothesis that videos with majority of frames to be faces, have a greater probability to gain bigger engagement like repost and likes, we explore these videos for ontological clues about what kind of description is prevalent amongst popular videos. For this analysis, we chose the popular image ontology and sentiment analysis method called DeepSentibank [5] which is a variant of Sentibank [4] by Boroth et.al, that uses deep convolutional neural networks [8] to train a classifier which classifies images into a set of 2089 classes. Each class essentially is a ontological description that contains one adjective and a noun which describes the image. This method gives a very accessible way to understand a higher level sentiment about a particular image. We use DeepSentibank

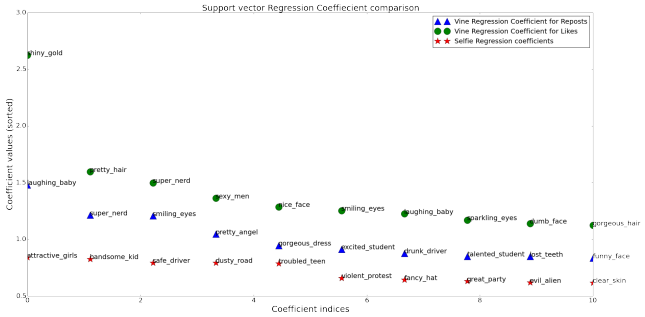

Figure 4: **Top 10 Support Vector regression coefficients for Instagram selfies , Vine Likes and Vine reposts**

on all the face frames extracted from videos. We then trained a simple Support vector regression on the dataset of face frames to understand if there is any correlation, and if there is which are the dominant ontological adjective noun classes. This exercise was performed for both Instagram selfies and video selfie frames. Interestingly, the vine frames a larger correlation of 0.32 for likes and 0.3 for reposts, as compared with Instagram selfies popularity of 0.2 correlation. We plotted the dominant regression coefficients for all the three in figure **??** to observe similar trends amongst all three. The adjective noun pairs however are different. Instagram selfies yield ANPs which are predominantly of positive sentiment, where as vine has a very mixed set. The higher correlation value shows a pattern might emerge at a higher abstraction using which prediction of reposts and likes for a video selfie vine could be possible.

7. PREDICTION OF LIKES AND REPOSTS IN VIDEO SELFIES

With the recent advances in software frameworks and tools that help in designing neural networks [2] [3] , training a neural network to learn a pattern in higher dimensional abstraction is easier. Because we found a decent correlation in the initial Support vector regression between Sentibank class probability distribution and repost/like distribution, we chose to use the 2089 dimensional ANP probability vectors that arise from Sentibank analysis as input random variable to our network. Before using them as is, we ran a Singular value decomposition on them to check if dimensionality reduction is possible or not. We found that a reduction to 1000 features from 2089 still captures 99% of the variance in data. Hence we reduce the feature vectors to 1000 dimensions using SVD reduction. Because the values of likes and reposts for a post are integer values, it was most important to quantise these values and convert the problem from regression to classification. The maximum like count on a video in our collected dataset was 92057 and the maximum repost count was 29084. So we divide like counts into 20 classes interval 5000 and we divide repost counts to 30 classes of interval 1000. This allowed us to convert the problem into a classification problem which would classify a video into one of the above interval class for reposts and likes. We converted all the likes and repost values of all the posts into one-hot vectors [6] which convert numbers into a sparse vector which has only the entry corresponding to membership class as 1 and everything else 0.

After all the pre-processing of the data, we build a simple 4 layered fully connected neural network with 1000 input neurons and 1000, 10000 ,1000 neurons in each of the subsequent layers as portrayed in figure 5. The final layer would have either 20 or 30 neurons based on what classification problem are we dealing with. We

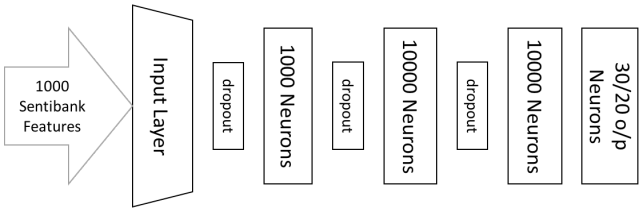

Figure 5: **Neural network architecture for vine repost and like predicton**

use ReLU as activation function and dropout layers to prevent overfitting as described by Dahl et.al [7]. We try to minimise the cost for categorical cross entropy between $P(Y)$ and $P^{\dagger}(Y)$, where $P(Y)$ is the true distribution of class label vectors and $P^{\dagger}(Y)$ is the predicted distribution of class labels for input random variable $P(X)$. So for any given training epoch the network is trying to minimize cross entropy H given by

$$H(y, y^{\dagger}) = -\sum_{Y} P(y) log(P^{\dagger}(y))$$

where y is the expected class vector for a given training sample and $y^{\dagger}$ is the predicted class vector by the network. The network was trained on 15000 extracted face frames from over 4000 qualified video selfies in a 3 fold cross validation fashion which resulted in 95% in sample accuracy for reposts and 98% in sample accuracy for likes at training time. We chose another 3500 face frames as out of sample validation set from different 1000 videos and got 90 % prediction accuracy for likes of those videos and 87.25% prediction accuracy for reposts.

Predictor	Out of Sample error	Correlation	MAE
Reposts	12.75 %	0.78	2.67
Likes	10.01 %	0.80	1.47

8. DISCUSSION AND FUTURE WORK

After this initial investigation, we can confidently say that like Instagram [1], vine and other video expression mediums would have a high engagement score for videos with faces. Further with the high correlation between face adjective noun pairs and repost or likes implies that there are certain kinds of sentiments in videos with faces, which invite much larger engagement from other users. As a future roadmap for this study, we would like to explore deeper into the sentiment and affective aspects of video selfies. The larger big picture ambition for this study is to delve into evoked sentiments using latent sentiments that touch upon in this paper. With the advancement of Deep learning tools, such work is on the horizon.

9. REFERENCES

[1] BAKHSHI, S., SHAMMA, D. A., AND GILBERT, E. 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* (New York, NY, USA, 2014), CHI '14, ACM, pp. 965–974.

[2] BASTIEN, F., LAMBLIN, P., PASCANU, R., BERGSTRA, J., GOODFELLOW, I. J., BERGERON, A., BOUCHARD, N., AND BENGIO, Y. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[3] BERGSTRA, J., BREULEUX, O., BASTIEN, F., LAMBLIN, P., PASCANU, R., DESJARDINS, G., TURIAN, J., WARDE-FARLEY, D., AND BENGIO, Y. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)* (June 2010). Oral Presentation.

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

[5] CHEN, T., BORTH, D., DARRELL, T., AND CHANG, S.-F. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586* (2014).

[6] COATES, A., AND NG, A. Y. The importance of encoding versus training with sparse coding and vector quantization. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (2011), pp. 921–928.

[7] DAHL, G. E., SAINATH, T. N., AND HINTON, G. E. Improving deep neural networks for lvcsr using rectified linear units and dropout. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (2013), IEEE, pp. 8609–8613.

[8] HINTON, G. E., SRIVASTAVA, N., KRIZHEVSKY, A., SUTSKEVER, I., AND SALAKHUTDINOV, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012).

[9] HU, Y., MANIKONDA, L., AND KAMBHAMPATI, S. What we instagram: A first analysis of instagram photo content and user types. In *ICWSM* (2014), AAAI.

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

[11] KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[12] SOUZA, F., DE LAS CASAS, D., FLORES, V., YOUN, S., CHA, M., QUERCIA, D., AND ALMEIDA, V. Dawn of the selfie era: The whos, wheres, and hows of selfies on Instagram. In *Proceedings of the 2015 ACM on Conference on Online Social Networks - COSN '15* (2015), pp. 221–231.

[13] VIOLA, P., AND JONES, M. J. Robust real-time face detection. *Int. J. Comput. Vision 57*, 2 (May 2004), 137–154.