Do sentiments tell a story: Exploring perceptual sentiments in high impact videos

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

1. ABSTRACT

This paper tries to explore the art of story telling in the realm of Online Social Networks (OSNs) and online social media. In the due course of the work done for the paper, we crawled a popular social media network called Vine for almost 2 months and collected over 12000 unique vine videos and their post meta data. We try to take an approach based on perceptual sentiment in social media and hypothesize existence of story lines in perceptual sentiments. We use deep learning tools to detect sentiment values of videos frames using the popular framework of Sentibank. The paper hypothesizes and shows to a reasonable extent, that perceptual sentiments do follow popular screenwriting theories. The paper also evaluates correlations of individual perceptual sentiments of videos with popularity metrics of the videos.

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 selfie videos because of the ubiquitous nature of cameras. A sharp changes in cultural aspects of online social interactions are evident and have also been studied in detail in papers like [14]. Naturally, properties of real life interactions are now going to emerge in virtual social interactions. A study by Daniel Perez et.al [12] observes that non-verbal communication has a strong effect on how a person is perceived at times of real world situations like interviews. These real world behaviours can also be extended to the virtual social world. With the rise of social media networks like Vine and Instagram, human to human non-verbal interactions have another dimension to manifest. The [14] paper explores several of these properties 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. These numbers are pretty much consistent amongst video OSNs like Vine, where in the protagonist of the video is in focus for most of the 7 seconds of the vine.

Taking into consideration these measurements from previous works, there seems an urgent need to bring these mediums under the overarching problem of sentiment analysis in social networks. That would allow the social network community to analyse and measure the dynamics of the world of social media in a different dimension

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

3. SENTIMENTS IN SOCIAL NETWORKS

Sentiments are fundamental part of our day to day social interactions. A face to face social interaction is generally augmented with facial expression, body language and linguistic sentiment to convey the exact meta information. These properties are very human in nature and are mimicked in the social networks as well. Studies like [8] have explored the world of linguistic sentiment in social networks, by comparing several popular textual sentiment analysis methods used for analysing tweets.

When it comes to social media shared over OSNs, the analysis becomes complicated. This is because social media involves higher dimensional messages like Videos, audio and Images. Moreover the media shared has a very human centric content. That means the media will involve a lot of faces, poses and affective means of communications. The studies done in [14] show that there has been 900 times increase in the number of selfies over Instagram in just 2 years. Another recent paper [9] states that everyday more than 90 million selfies are taken using just the Android clients out there and are uploaded on Instagram. We collected Vine social network data, which is a popular social network that uses short 6 seconds videos as a medium. In that dataset we found that one in ever three video in the popular videos category contain human faces for more than 60 percent of the frame length. The very human centric nature of the media shared over these networks, make sentiments and human affects an integral part. These mediums When it comes to perceptual sentiments, there are two broad categories that could be explored. The first category looks at the perceptual sentiment evoked by a social media content. The second category talks about the actual latent perceptual sentiment that comes with the context of the content itself. We will discuss about the research problems about both these categories.

3.1 Evoked perceptual sentiment

Several works have done in depth studies using methods like crowdsourcing to understand the different shades of a particular evoked emotion. Works like UrbanGems [1] and StreetScore [13] use crowdsourcing methods to understand degrees of human sentiment evoked because of pictures of real urban neighbourhoods. Sentiments like the feeling of safety and aesthetics are especially

hard to quantify and crowdsourcing helps the authors to do some interesting modelling. On the other hand there are papers like [7] by L. Jeni et.al. describe utility of actual facial expression detection for understanding content consumer reaction. Such approaches help us understand the very effect of a particular content on the consumer.

3.2 Latent perceptual sentiment

This approach is what this paper stresses on. By latent perception, we mean the hidden parameters, which are part of the very content. Social networks like reddit have specific sub-reddits that work on appealing to these types media sentiments that evoke emotions like empathy, disgust, contempt and love . One such popular sub-reddit is labelled R/aww which contains images and GIFs that showcase cute animals and animal behaviours. Another one called R/cringe appeals to the sentiments of awkwardness and discomfort by exhibiting videos and Gifs about people in awkward situations. These specific social channels are popular because the content shared over these channels have a certain type of latent sentimental response, which the consumers of these channels resonate with.

Our paper focuses on this part of the story, and tries to survey and benchmark certain state of the art methodologies out there. We also propose certain hybrid approaches, which show that we can attain much better performance if a heuristic approach to combine certain methods is taken.

4. SENTIMENT ANALYSIS METHODS

To the best of our knowledge we have evaluated certain popular approaches in solving the problem of extracting latent sentiment in a media content. The sentiment analysis methods broadly fall into two bins. One is the Content based Image retrieval (CBIR) [11] set of approaches, which actually analyse the image structure and contents to extract features and inferences about the image. The second bin is emotional semantic image retrieval (ESIR) [15] which aim at trying to extract the semantic gist of a particular image. Human brain is great at extracting such semantics. For example it is very natural for a person to describe a particular image as "picturesque" or "scenic" or to describe someone's clothing as "tacky" , "classy" or "elegant". These semantic classes, no matter how subjective, are also sufficiently descriptive for another human being to process. In the subsections to come, we will discuss some of the popular perceptual sentiment analysis methods.

4.1 Facial Action Coding System (FACS) based methods:

Facial Action Coding System (FACS) based approach towards understanding human affects was the pioneering research done at CMU that paved the way of modern affective computing. The paper [10] talks about this pioneering research. The method works on a very important base of Facial Action Coding [4] system which encodes movement of specific muscles of human face and encodes actions units as a combinations of one or more of these movements. These AUs or Action Units are then carefully measured from frontal images of faces and then models are built on top of these measurements to classify different emotions.

The paper evaluates these AU based methods for specific AUs. They do not show a very convincing performance analysis for the 7 prime emotion bins, but evaluate classification accuracy for each of the AUs. The performance shows very promising accuracy for detecting AUs in upper and lower part of the face. Works like [2] [3] take these concepts forward and use features in AU zones to build classifiers. These studies employ features like Pyramids of

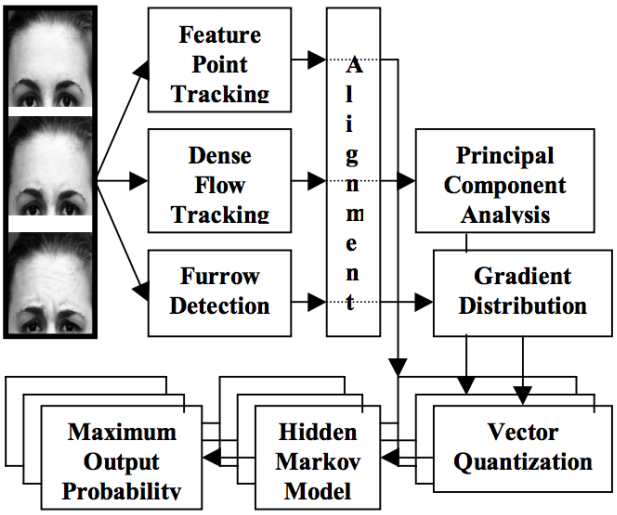

Figure 1: **System architecture of FACS AU based Emotion recognition**

Histograms of Gradients (PHOG) and Local Phase Quantisation (LPQ) to classify across 5 key emotions of anger, fear, joy, relief and sadness. Their methods attain a very impressive performance bracket of 67 to 74 percent detection accuracy.

4.2 Semi-supervised learning models

Over the past two decades, with the surprising developments parallel computing and general purpose graphics processor computing, the space for scaling up and parallelising sparse computation has increased exponentially. This paved the way for extremely fast and scalable neural network frameworks. With these frameworks, implementing complicated network designs became possible.

One of the researched problem was in the area of affective computing and emotion recognition. A very peculiar type of network which was put to test to solve this problem was the Deep Belief Network [6] . Deep belief networks come under the category of semi-supervised learning, where in the network tries to extract a higher level representation just by operating on the statistical properties of the input data, and then the network is further fine-tuned using the supervised labels. Such a popular approach was first demonstrated by Hinton et.al [6] using something called as Deep Belief networks which are basically stacked restricted Boltzmann machines. But as shown by Hinton in [5] , these networks are really hard to train in a supervised manner unless the individual weights are brought to optimal values before the network is trained in supervised manner.

4.3 Supervised Learning Models

As mentioned earlier, the problem space for machine learning can be broadly fragmented into two wide categories. When the target function to be optimised has a predefined class structure, it is called supervised learning models. The advantage of this method is that it optimises for a very well defined target function. The drawback however means that you need large amounts of reliable and labelled data. For this very problem of affective classification, we do happen to have certain datasets that we could use for training a network model.

5. REFERENCES

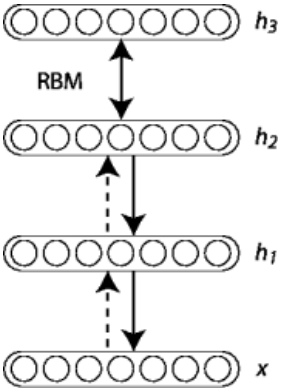

Figure 2: **Deep belief networks as Stacked Restricted Boltzmann machines**

[1] ADAM BARWELL, DANIELE QUERCIA, J. C. http://www.cam.ac.uk/research/news/how-to crowdsource-your-happy-space, 2012.

[2] BUSSO, C., DENG, Z., YILDIRIM, S., BULUT, M., LEE, C. M., KAZEMZADEH, A., LEE, S., NEUMANN, U., AND NARAYANAN, S. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th International Conference on Multimodal Interfaces* (New York, NY, USA, 2004), ICMI '04, ACM, pp. 205–211.

[3] DHALL, A., ASTHANA, A., GOECKE, R., AND GEDEON, T. Emotion recognition using phog and lpq features. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on* (March 2011), pp. 878–883.

[4] EKMAN, P., AND FRIESEN, W. V. The facial action coding system. *Consulting Psychologists Press Inc. San Francisco CA* (1978).

[5] HINTON, G., AND SALAKHUTDINOV, R. Reducing the dimensionality of data with neural networks. *Science*, 5786 (July 2006), 504 – 507.

[6] HINTON, G. E., OSINDERO, S., AND TEH, Y.-W. A fast learning algorithm for deep belief nets. *Neural Comput. 18*, 7 (July 2006), 1527–1554.

[7] JENI, L. A., LŐRINCZ, A., NAGY, T., PALOTAI, Z., SEBŐK, J., SZABÓ, Z., AND TAKÁCS, D. 3d shape estimation in video sequences provides high precision evaluation of facial expressions. *Image and Vision Computing 30*, 10 (2012), 785–795.

[8] JOO, J., LI, W., STEEN, F. F., AND ZHU, S. C. Visual persuasion: Inferring communicative intents of images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2014), pp. 216–223.

[9] 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.

[10] LIEN, J. J., KANADE, T., COHN, J. F., AND LI, C.-C. Automated facial expression recognition based on facs action units. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on* (Apr 1998), pp. 390–395.

[11] LIU, Y., ZHANG, D., LU, G., AND MA, W.-Y. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition 40*, 1 (2007), 262 – 282.

[12] MARCOS-RAMIRO, A., PIZARRO, D., MARRON-ROMERA, M., AND GATICA-PEREZ, D. Let your body speak: Communicative cue extraction on natural interaction using rgbd data. *IEEE Transactions on Multimedia 17*, 10 (Oct 2015), 1721–1732.

[13] NAIK, N., PHILIPOOM, J., RASKAR, R., AND HIDALGO, C. Streetscore – predicting the perceived safety of one million streetscapes. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on* (June 2014), pp. 793–799.

[14] 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.

[15] WANG, W., AND HE, Q. A survey on emotional semantic image retrieval. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on* (Oct 2008), pp. 117–120.