Automatic User-Video Metrics Creations From Emotion Detection

Darari Nur Amali[1], Adnan Rachmat Anom Besari[2], Ali Ridho Barakbah[3], Dias Agata[4]
Department of Informatics and Computer Engineering
Politeknik Elektronika Negeri Surabaya (PENS)
Kampus PENS, Jalan Raya ITS Sukolilo, Surabaya, 60111 Indonesia
Email: [1]dararii@ce.student.pens.ac.id, {[2]anom, [3]ridho, [4]diasagata}@pens.ac.id

***Abstract* - In this digital era, digital content especially video, is increasing in number from time to time. Typically, a video service provider like Youtube will perform video analysis based on the video content such as colours, textures, shapes, and other features that exist in video content. The result of this analysis was used to understand user preference and to personalize video for each user. With technological developments, especially in Machine Learning and Computer Vision technology, video analysis can be based on other things beyond the video. In this context, it is the audience's impression. Thus, with the analysis of audience impressions in real-time, it is expected that the video can be analysed using the emotion parameters of the audience while the video is playing, and this can be done automatically and real-time. This system generates impression statistic for each video which concluded from every user who has watched the video and save those data in the database. Method used to analyse the result is by recruiting respondent and give some questionnaires. Respondents were asked to watch some videos and were asked to compare the impression metric which created by the system with user's real impression. The result shos that the automatic video-metric creation from emotion detection has been able to measure user's impression of the video with more than 80% accuracy stated by 75% of 20 respondents of the survey.**

Keywords — Video Labelling, User-Video Metric Creations, Emotion Detection.

I. INTRODUCTION

Communications network infrastructure in Indonesia develops rapidly in recent years. Survey of the Indonesian Internet Network Provider (APJII) in 2016 revealed that 132.7 million Indonesians are connected to the Internet [1]. This has led to an increase in the use of digital technology-based facilities in the community, one of them is video streaming platform. This triggers streaming video service providers such as Youtube increasingly preferred by many people.

Youtube-statistic reported that there are 300 hours of videos uploaded to Youtube (refers to Fig.1). The large number of videos on a streaming service such as Youtube will make the video on the service very hard to categorized. This will encourage providers to analyse the video content through the internal content of the video or video features. The features are colour, shape, texture, audio signal model, and others [2]. Then they used the result to determine a video category, get user preference in watching video, and create a personalization for each user.

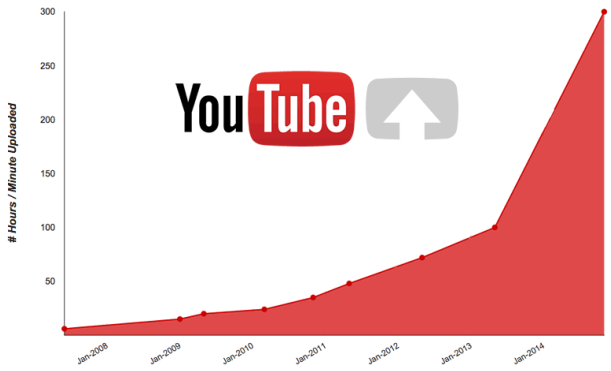
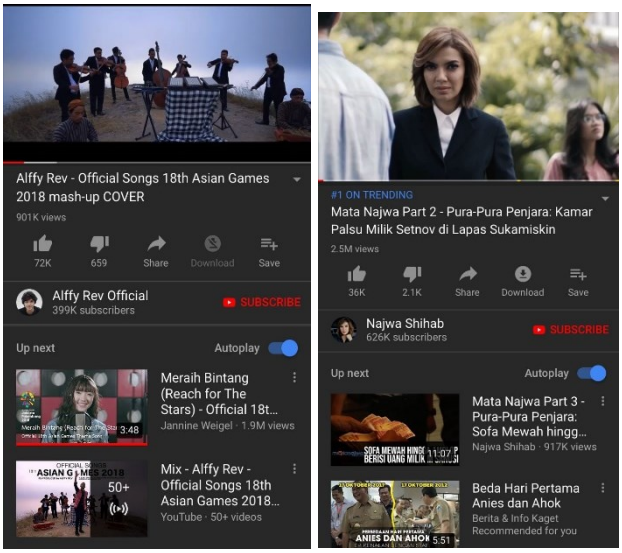


Fig. 1. Statistic of video uploaded to Youtube in every hour since January 2008 to January 2014.

Their algorithm in analysing video content through its feature was very powerful, with billion videos in numbers as data set, the categorization of the video performed very great in accuracy. So, Youtube can provide many videos with same category as video which played by user, and this system is a part of their recommendation system. But, there are some weaknesses, video which provided based on same feature of its video can't provide variety videos for user, and in some cases, user be bored with the video provided by Youtube recommendation system. In fig. 2 (a), a first case that music video played by user, and Youtube recommend another music video for user. In fig. 2 (b), a second case, reality show video was played by user, and Youtube provide another reality show video. The weakness is, when user wants a video with different category or different content, user cannot find it at video list provided by Youtube and should use search function to get them.

With development of massive artificial intelligence technology, video analysis can now be done through external factors that exist beyond its video, in this case is the impression of the video audience while watching the video. This system is designed to be able to auto-label for each video that is in accordance with the user's impression. Then, after the impression data of each video is obtained, then we will be able to find out whether the video makes the audience laugh, surprised, sad, fear, or even disgust. This will be automated by system and work in real-time. The goal of this research is to create an impression-based label for each video that exist and stored in a database. The result of an impression analysis can be used as a parameter and data set for a new recommender system in further research.

978-1-5386-8402-3/18/$31.00 ©2018 IEEE



(a) (b)

Fig. 2. (a) music video recommendation list, (b) reality show video recommendation list

II. METHOD

A. *Emotion Detection using Affectiva*

Affectiva is one of the projects of the Massachusetts Institute of Technology (MIT) that focuses on machine learning and computer vision. Affectiva Emotion API can perform the scan of a person's facial expression in real-time [3]. In the process of detecting emotions through facial images, Affectiva has 4 stages: 1) Face and facial landmark detection, 2) face texture feature extraction, 3) facial action classification and 4) emotion expression modelling [4]. These four stages can be seen in the Fig.3. This system requires an image as an input which can be obtained from camera stream, video file, video frame stream, or even a picture. To obtain image using camera stream, Affectiva needs camera with minimum specification among them:

1. RGB Camera
2. Minimal Resolution at 320 x 240
3. Minimum video capture speed at 10fps.

This algorithm can run at the background and computational resources which is used to determine an emotion varied based on processing speed. Usually between 1 to 10 FPS (Frame per Second).

As the algorithm shown in Fig.3, Affectiva is able to detect emotions well with accuracy score varied between 0.72 up to 0.9 at ROC (Receiver Operating Characteristic) curve [3]. Here is an example of test data that has been done to get some facial action (Fig. 4).

Facial action describes facial expression of human face. To understand the emotion, Affectiva needs to analyse some facial action that occurs then determine an emotion. Affectiva will generate score 0-100 in determining the emotional score of a person's facial expression. This score determined based on the EMFACS mapping algorithm developed by Friesen & Ekman. The following table I shows the mapping of facial expressions to emotional expressions.

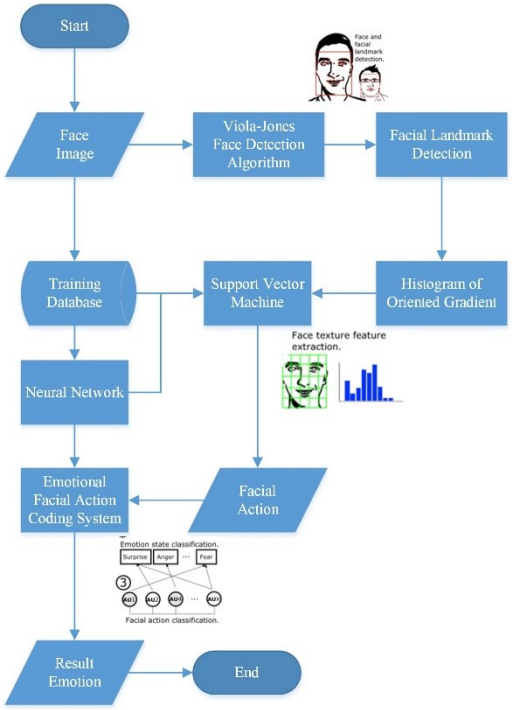


Fig. 3. Emotion detection system flow which developed by Affectiva [4]

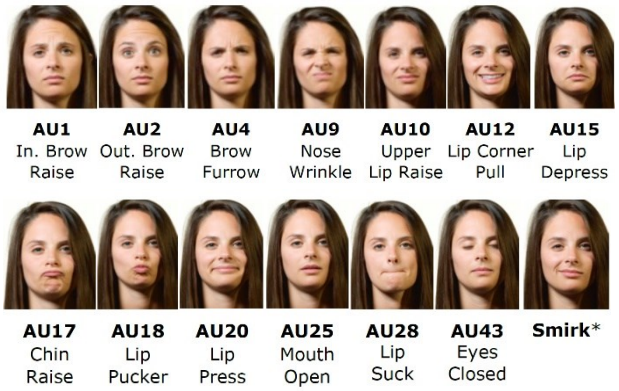


Fig. 4. The facial actions that can be detected. Each action is given a score from 0 to 100. [4]

TABLE I. Expression Mapping to Emotion

Emotion	Increase Likelihood	Decrease Likelihood
Joy	• Smile	• Brow Raise • Brow Furrow
Surprise	• Inner Brow Raise • Brow Raise • Eye Widen • Jaw Drop	• Brow Furrow
Disgust	• Nose Wrinkle • Upper Lip Raise	• Lip Suck • Smile
Fear	• Inner Brow Raise • Brow Furrow • Eye Widen • Lip Stretch	• Brow Raise • Lip Corner Depressor • Jaw Drop • Smile
Sadness	• Inner Brow Raise • Brow Furrow • Lip Corner Depressor	• Brow Raise • Eye Widen • Lip Press • Mouth Open • Lip Suck • Smile

B. Combine Emotion Detector with Streamer

The automatic video-metric creations system will be implemented on Android Apps. The apps will develop using Android Studio and Affectiva SDK for Android. Android is chosen because Android OS market share in sales to end users from 1st quarter 2009 to 1st quarter 2018 achieved more than 85% in worldwide [5].

The flow system of this metric will be based on the flowchart in fig. 5 it can be seen that Youtube video will run simultaneously with emotion detection system. The metrics of the user-based impression video will be counted every second. So, the metric results will appear discrete with 1 second interval. Analytics from user-video metrics will be done to determine the impression conclusions on a video.

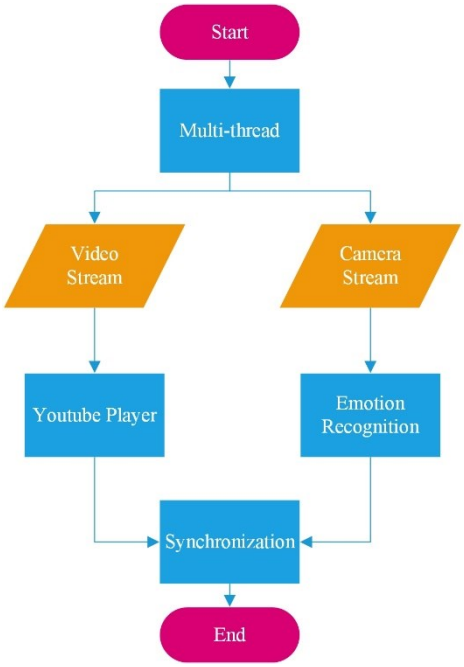


Fig. 5. General system flow of automatic video-metric creations.

Impression details will be stored in databases that have been designed for efficiency and analysed by the system to get an impression of the conclusion of the total impressions of existing video and stored as meta-data that will facilitate the system in accessing data efficiently and quickly [6]. The data to be formed is shown in fig. 5.

TABLE II. Metric data shape shown in table

Video-Metric	V1	V2	....	V(n-1)	Vn
User 1	X 1.1	X 1.2	X...	X 1.(n-1)	X 1.(n)
User 2	X 2.1	X 2.2	X...	X 2.(n-1)	X 2.(n)
.....	X...	X...	X...	X...	X...
User (n-1)	X (n-1).1	X (n-1).2	X...	X (n-1).(n-1)	X (n-1).(n)
User n	X (n).1	X (n).2	X...	X (n).(n-1)	X (n).(n)

Database structure shown in table II was only meta-data structure. The database design will use flat design structure using Firebase Firestore platform. Firebase Firestore was chosen because its speed and reliability. So, the table model design shown in table II will be converted to flat database design shown in Fig. 6.

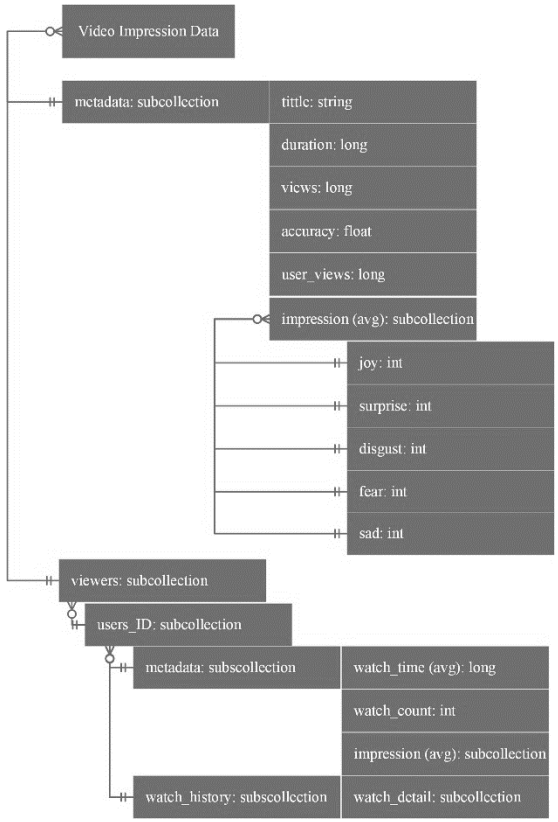


Fig. 6. Video-metric database structure shown in table 2 converted to flat design database which supported by Firebase Firestore.

Flat data structures are chosen because non-SQL data structures have much faster speed compared to table-shaped SQL in very large data conditions (more than 10,000 data). With the above structure, the database will be scalable. So, this database will not be limited by how much the amount of video impression data will be stored later. The database structure for users is also designed scalable, fast and efficient. In this case, using the meta-data technique. Meta-data is used to perform indexing on video data. This is intended to avoid duplication of data so that the storage space used is also efficient.

The value of X in table II is obtained by calculating the accumulation of emotion value which is done by algorithm in fig. 7. The X value is five-dimensional data which contain $X_{[i=1]}$ to $X_{[i=5]}$ where i is the five emotions category such as joy, surprised, disgust, fear, and sad.

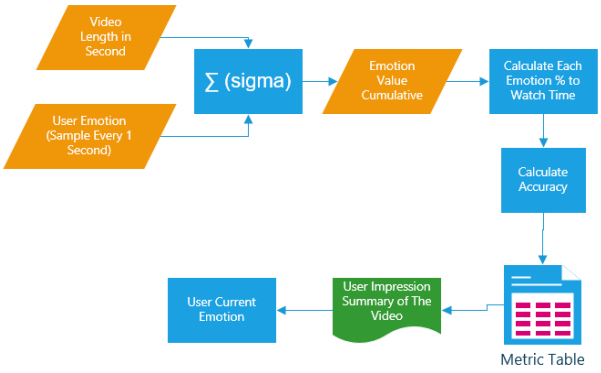


Fig. 7. User-video metric calculation and data storage algorithm

The algorithm in fig. 7. shows that the value to be stored in the metric table is the value of six emotions except the neutral emotion. Value calculations can be done as this equation.

$$\Sigma\ Emotion\ Value = \sum_{k}^{n} (Emotion\ Value[k]) \quad (1)$$

The calculation of equation (1) will get the accumulative emotional value. Where n is the amount of emotional data (ideally equal to the video duration in seconds). Emotion Value [k] is the viewer's emotional value on the video on the duration of k. This accumulative value will be compared with the accumulated value of maximum emotions score. So, that it will form the value of emotion percentage to watch time in accordance with calculations of equation 2.

$$Emotion\ Metric\ \% = \frac{\Sigma\ Emotion\ Value}{\Sigma\ Max\ Possible\ Emotion\ Value} * 100\% \quad (2)$$

The calculation for the maximum emotional value is obtained from the accumulated calculation of the maximum emotional value during the duration of the video. This calculation is in accordance with the following equation 3.

$$\Sigma\ Max\ Possible\ Emotion\ Value = \sum_{t}^{n} (Max\ Value[t]) \quad (3)$$

Where n is the duration of the video, and t is the index of time of the video. So, max value [t] is the maximum value of the emotion to be detected on the duration of t. Maximum value on this system is 100.

In an automated video-metric manufacturing system, the measured data must meet the requirements to be stored in the database. The requirements are metric data which must have user impression data minimum 70% of video duration. This is intended to maintain the level of measurement accuracy. To measure the accuracy level of this video metric, the system has been able to perform automatically that is by comparing the number of detected expressions with the total duration of the video according to equation 4 below.

$$Metric\ Accuracy\ \% = \frac{\Sigma\ Emotion\ Detected}{\Sigma\ Video\ Duration} * 100\% \quad (4)$$

According to equation 4, total count of emotion detected is compared to video duration to calculate Metric Accuracy. This is because system use 1 frame per second in processing rate. So, there will be one emotion value at every second. At 100% accuracy, the total count of emotion value will be the same with the video duration.

III. Result and Analysis

This test is intended to discover whether the system has been able to display a video gauge based on the audience's impression when viewing it. The test examined several videos with diverse users.

A. Validation Method

Validation method for the test result is done by comparing the emotion result calculated by system with the impression characteristic of the video which estimated manually by the authors. If the first step of the validation is valid, then the system can validate the video label by itself, if the video has been watched in 10 times by some users. The last validation step is by taking feedback from users using questionnaire. In the questionnaire, users will asked to compare the emotion metric statistic which has been created by system to their real emotion feeling while watching the video.

B. Test I.

Video Title: Djarum Video Ad – Want to be skinny
Video Duration: 32 Seconds
Video Category: Joy Video/Comedy
Video Characteristics: This video is a comedy video ad such as a video ad of PT. Djarum in general, but the location of this cuteness is located at the end of the video.
Audience Characteristics:

- Age: 20 - 25 years
- Sex: Male and Female
- Mood: Unknown (Unknown)

In this test, it be seen how the user's impression of this video. Video metrics are said to be true if the user's impression of a video has something in common with the video characteristics tested.

From the result of this test, it shows that the total happy impression reached 37.43% which is dominated in the 23rd to 32nd position in this video. Joy impression at this result test also can be seen at 6th to 12nd position referred to fig. 8. This is in accordance with the characteristics of the video being tested. Up to this point, based on table III, impression-based video measurement is considered correct.

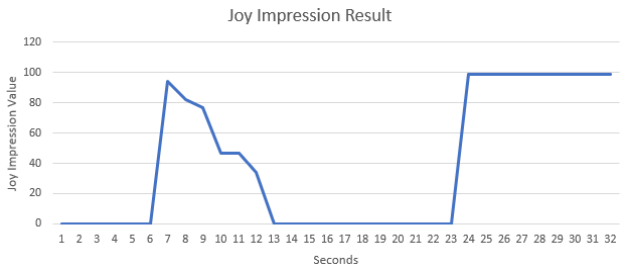


Fig. 8. Users impression detail result of video 1 by user 1 created by automatic video-metric system

TABLE III. Video 1 Impressions Result by 10 Users

No.	Video ID	User ID	% Joy Value
1	b8vhQ-	2zg0xsiPXqZaV8S6SBCboVPHDo23	11,12
2	b8vhQ-	ArL6uaqHvgc3XPlvPCiK40KOJv53	27,25
3	b8vhQ-	bNCD6YWBlkN6SN6xm4CRzKIOF6n1	8,27
4	b8vhQ-	djasMlryLZemk9SWo9KvGRkvJfg1	39,33
5	b8vhQ-	iEnojOVPPwY75GhC2xGhcrDGVgX2	4,27
6	b8vhQ-	jMF5BlXUUgS79q7byH2yS1Fnpg52	6,26
7	b8vhQ-	P33JdE8VVoejEvwWnS246l5Gf3H2	9,31
8	b8vhQ-	St8Xq66JjOWRRqSKVw0zqv6NkeC3	10,92
9	b8vhQ-	YDC5vvkOIsO6YWQEoBBsPYdf0vJ3	11,61
10	b8vhQ-	zLzIWuam2jWn2Kzz1ST1qzZU7qb2	72,47

C. Test II.

Video Title: Asian women eating frog
Video Duration: 59 Seconds
Video Category: Video Log – Disgusting for some people

Video Characteristics: This video is a video documentation of a female vlogger who is eating frogs. For most people, frogs are a disgusting animal to eat, so this video is suitable for use as a test for disgusting videos.

From the results of tests conducted for disgusting videos involving 7 users in 10 times views, the results show that this video has an abhorrent label on average of 25% of the expression of disgust detected when users watch the video in accordance with the experimental results listed on the table IV.

TABLE IV. Video 2 Impressions Result by 10 Users

No.	Video ID	User ID	% Disgust Value
1	iDoA9v	kZBvxp7Vx1Q7aYdei3gbk1UePZg1	45,82
2	iDoA9v	dta7oYAJlaZST9CU9Cye0RSb8x13	13,59
3	iDoA9v	dW5ix8hGLYQcgBG2mye5WUP4QUi	19,34
4	iDoA9v	T2MoWeBgu0UMBxOHTiHeteIEDsu2	0,41
5	iDoA9v	dta7oYAJlaZST9CU9Cye0RSb8x13	41,81
6	iDoA9v	S0VjmLrdQgNWQNZ0VcXniee6eH92	14,99
7	iDoA9v	P33JdE8VVoejEvwWnS246l5Gf3H2	20,58
8	iDoA9v	9sKjAHYnTjNA9AUHUDQzDaCJYb	18,16
9	iDoA9v	P33JdE8VVoejEvwWnS246l5Gf3H2	75,02
10	iDoA9v	P33JdE8VVoejEvwWnS246l5Gf3H2	0,47

In one user's experiment, disgusted impression was detected at 74.14% with a disguised impression spread evenly at 0th to 50th seconds. This appears as shown in Fig. 9. Based on the results of this test, the system has successfully detected a disgusted impression on the disgust category video.

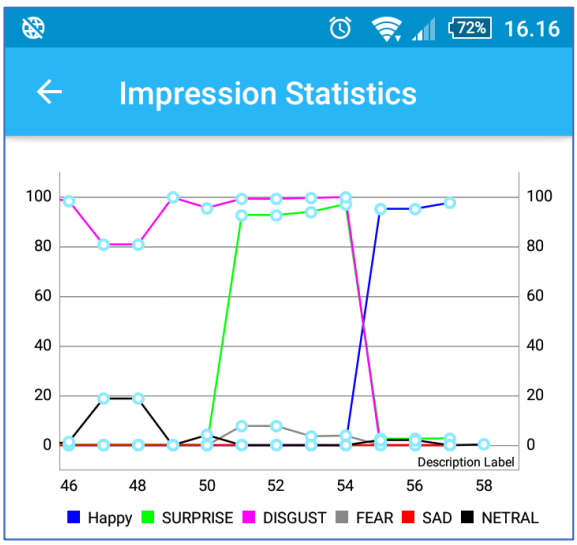


Fig. 9. Users impression detail result of video 2 by user 3 created by automatic video-metric system

D. Test III.

Video Title: Man Sleep on Rail when Train is Passing.
Video Duration: 52 Seconds
Video Category: Amateur/Surprising video at first watch.
Video Characteristics: This video is an amateur video that is the documentation of a man who is doing a test of guts by sleeping in a railroad when the train is passing. The surprising thing about this video is that it does not seem at first that there will be a passing train, but suddenly a train passes by the 18th second and the man goes straight to sleep on the tracks as the train passes.

From the results of the tests conducted for a surprising video involving 7 users in 10 views, the results show that this video has an astounding average label of 7% of the surprised expressions detected when the user watches the video in accordance with the experimental results listed on the table V.

TABLE V. Video 3 Impressions Result by 10 Users

No	ID	User ID	% Surprised Value
1	aZlA7	2gBcgXWHWgeLvSRs7HwPjIF5	14,83
2	aZlA7	9sKjAHYnTjNA9AUHUDQzDa	4,04
3	aZlA7	HnjBYINrKcPlWb9dsXba0BTSmQ	11,47
4	aZlA7	P33JdE8VVoejEvwWnS246l5Gf3	2,33
5	aZlA7	dW5ix8hGLYQcgBG2mye5WUP	4,17
6	aZlA7	P33JdE8VVoejEvwWnS246l5Gf3	15,78
7	aZlA7	dta7oYAJlaZST9CU9Cye0RSb8x	3,87
8	aZlA7	P33JdE8VVoejEvwWnS246l5Gf3	2,05
9	aZlA7	dta7oYAJlaZST9CU9Cye0RSb8x	7,09
10	aZlA7	kZBvxp7Vx1Q7aYdei3gbk1UePZ	4,81

In the experiment for user 1, the surprised impression can be seen at the 18th to the 23rd seconds (in figure 10), this is because at that moment, the video shows an incident of a man sleeping directly on the train track as the train passes. And this can make the audience surprised in seeing the action. The surprised expression was also seen in the 44th minute to the end of the video, because at that moment the video showed that the man was not hurt even though he slept on the train track as the train passed. It can also make the audience surprised. For the results of experiments on one of the users, can be seen in the following Fig. 10.

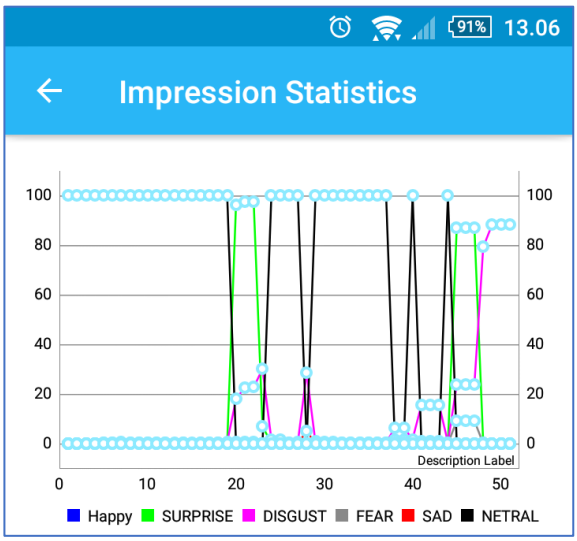


Fig. 10. Users impression detail result of video 3 by user 4 created by automatic video-metric system

E. Accuracy

The last step of result validation is by getting feedback from users which have watched the videos using this application. By using questionnaire method, users are asked to compare the impression metric which have been created by system with their real feelings. In this research, feedbacks are collected from 20 respondents.

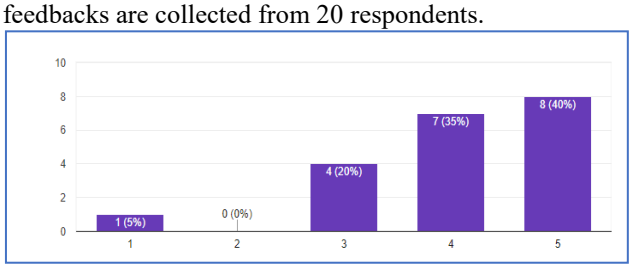


Fig. 11. Survey result from users' real feelings from 20 respondents

Respondents are asked to choose scale 1 – 5 corresponding to measurement accuracy. The scale 1-5 shows measurement accuracy of 20%, 40%, 60%, 80%, and 100% respectively. The accuracy of automatic video-metric measurement system from the results of user's feedback (refers to Fig.11) can be seen that 75% of respondents stated accuracy exceeds 80%, 20% of respondents said the accuracy is only about 60%, while one respondent stated accuracy of less than 20%.

IV. CONCLUSION

The automatic user-video metric creations system has been able to measure user impressions with an accuracy of more than 80%. This level of accuracy is expressed by 75% of the 20 respondents during the test. In addition, it is known that the ideal distance the system needs to detect the emotions of a user's facial expressions is 20-40 cm from the front of the user's face.

From the results of user impression testing of video using a particular video on different users, the system produces different impressions result. As for things that affect the results of the measurement of the impression is the position of the smartphone to the user, the user's behaviour during watching the video, racial differences between the developer of emotion detection system with the respondents on the test, the difference of the category perception video tested, and the psychological condition of the respondent's influence during the test. In general, automated user-video metric creations systems have been able to work properly, with this result, this system can be used as prerequisite to create an impression-based model video recommender system in the future research.

ACKNOWLEDGMENT

This research was partially supported by RISTEK DIKTI through Student Creativity Program (PKM). We thank our colleagues from Politeknik Elektronika Negeri Surabaya who provided insight and expertise that greatly assisted and testing the research results, although they may not agree with all of the interpretations/conclusions of this paper.

REFERENCES

[1] APJII Infographic Survei Presentation 2016. [Online]. Available: https://apjii.or.id/downfile/downloadsurvei/infografis_apjii.pdf. [Accessed: 30-April- 2018].

[2] Covington, Paul. Jay Adams. Emre Sargin "*Deep Neural Networks for YouTube Recommendations*". Google Mountain View, CA. 2016.

[3] D. McDuff, R. el Kaliouby, T. Senechal, M. Amr, J. F. Cohn and R. Picard. *Affectiva-MIT Facial Expression Dataset (AM-FED): Naturalistic and Spontaneous Facial Expressions Collected In-the-Wild*. IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, 2013, pp. 881-888. 2013.

[4] McDuff, Daniel & Mahmoud, Abdelrahman & Mavadati, Mohammad & Amr, May & Turcot, Jay & el Kaliouby, Rana. *AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit.* (2016).

[5] Statista. *Global mobile OS market share in sales to end users from 1st quarter 2009 to 1st quarter 2018*. 2018. [Online]. Available: https://www.statista.com/statistics/266136/global-market-share-held-by-smartphone-operating-systems/. [Accessed: 30- May-2018].

[6] Maksym Petrenko, Amyris Rada, Garrett Fitzsimons, Enda McCallig, Calisto Zuzarte. *Best Practices Physical database design for data warehouse environments*. 2012.

[7] C. Győrödi, R. Győrödi, G. Pecherle and A. Olah, "A comparative study: MongoDB vs. MySQL," 2015 13th International Conference on Engineering of Modern Electric Systems (EMES), Oradea, 2015, pp. 1-6.

[8] Y. Li and S. Manoharan, "A performance comparison of SQL and NoSQL databases," 2013 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), Victoria, BC, 2013, pp. 15-19.

[9] McDuff, D.J., Mahmoud, A.N., Mavadati, M., Amr, M., Turcot, J., & Kaliouby, R.E. (2016). *AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit*. CHI Extended Abstracts.

[10] Rolfe Winton, Rana El Kaliouby. *Measuring Emotions Through A Mobile Device Across Borders, Ages, Genders and More*. ESOMAR. 2012