Relation between facial expression and user described episodic memory timeline

HASEGAWA SHOICHI[†1] LOPEZ ERIK[†1]

Abstract: Narrative or episodic memory is a way of how humans recall a story. Narrative has also a strong connection with human emotions and attention. We combine the analysis of face expressions and attention to find a way to generate narrative timelines that match each individual person. We use face shape deformation analysis and eye gaze tracking to understand when a person has a change of facial expression and associate this with a narrative event. To evaluate our assumption, we performed an experiment where several subjects watch a short video and play a game while we recorded their physical reactions. Then, our software generate a narrative timeline based on those reactions. We conclude that this narrative timeline is a close representation of what the user considered the most memorable events of the experienced content.

Keywords: Narrative, episodic memory, face expression, attention, face deformation, eye gaze

1. Introduction

Human-computer interaction has increased in the last decade with the exponential growth of technology and software development.

A narrative is not only a story but also a series of memorable events that a person can remember when experiencing a certain situation. By definition, a narrative has a beginning, a middle and an end. From the semiotic perspective, narrative would be a process of reordering and telling again, what has already happened [1]. According to the Greek philosopher Aristotle [2], humans, by nature, can use stories to learn and experience new things. This experience plays an important factor when creating narrative.

Video games are not films, adventure novels or plays by themselves, but they share traits with other forms of cultural production like books and movies. In addition, not all video games tell stories, but many of them have narrative aspirations. In many actual games, we can clearly see the game designer's intentions to create a series of narrative experiences for the players to enjoy. Then, we consider the game designers as narrative architects. [3]

Through video games, the players can experiment diverse situations. If, during the gameplay, a specific situation has enough emotional impact, these should generate a physical reaction on the player and these experiences should generate an emotional reaction that the players should be able to remember. Therefore, while playing video games or while looking to video gameplay, the individuals build their own individual emotional narrative, depending on which gameplay situations affect them. [4][5].

According to Salen and Zimmerman [6], game design is the process by which a game designer creates a game, from which meaningful play can emerge; therefore, the goal of a successful game design is the creation of meaningful play. The descriptive definition of meaningful gameplay is the relation between the player's action and the outcome of it. The evaluative definition for meaningful gameplay involves the immediate outcome that a player perceives and the impact that this outcome has in the game as a whole. For our research, we need to analyze meaningful gameplay, and for that, we need to ensure the player through a narrative timeline can recall the game.

2. Related Work

We analyzed previous neuroscience research that give support to our theory that an emotional activation causes an impact on memory. Cahill [4] mentions that: "It is well known that emotionally arousing events are more likely to be encoded into long-term memory as compared to neutral events". Furthermore, Dere [5] hypothesizes that the emotional activation seems to be a requisite to trigger episodic memory formation. Therefore, we can certainly assume that an emotional impact generates an episodic memory formation.

There are several previous projects focused on measuring the users' experience while watching movies. For example, Chu's [7] proposal is very similar to ours. In Chu's research, they measure user's facial expression and eye movement variations while watching several videos. Then they analyzed and synchronized the measured data with the video, to summarize automatically the movies and reduce efforts of manual video summarizations. Chu only measures the user's physical response to generate a video summary, whereas we focus on analyze the user's emotive experience by comparing the user's recalls and physical reactions while watching the video.

3. Proposed Method

We define the hypothesis of this paper stating that there is a way to find what remains in a user's memory based on facial expressions produced by a narrative. As we mentioned before, humans can gain experience and sense about the real world though storytelling [1]. With this idea in mind, it becomes of a great importance to know how to develop a memorable narrative through a good experience.

From this point, it is necessary to find a way to understand the user's narrative. We base the approach of this paper on tracking the person's attention and facial expression; if we can know how the person is reacting to a specific element, then we can make a

[†1] Tokyo Institute of Technology

relation between a memorable event and a face expression.

3.1 Narrative Analyzer Software

The approach for analyzing the narrative is developing a software to find the factors that are relevant to the user, such as characters, items and so on.

The core functionality of this software is to be able to recognize each time the video motivates the user, leading to a memorable event ready for adding to the narrative timeline. We need to find a way to know when the user had a memorable event. One way to accomplish this, is analyzing physical reactions to the video. In this specific software, we used two sensors for tracking the user reactions based on the eye gaze and face.

For the eye gaze we use the Tobii EyeX sensor, which can tell us the area of the screen where the user is looking at. The eye gaze is essential to determine which object on the screen will be recalled by the user and added to the narrative timeline.

For the face, we use the Microsoft Kinect 2 sensor, which can give us, in high detail, the user's face shape deformations. Microsoft Kinect 2 SDK provides 70 face shape deformation indexes and 17-shape-animation indexes from which six are animation units and 11 are shape units. The shape units or SU's weights indicate how the face shape differs from the average shape. The animation units or AU's are variations from the neutral shape and provides us with details of the facial expression.

3.2 Experiment Workflow

A session on our Narrative Analyzer software works as follows:

Step 1. Calibration

It is necessary to carry out a calibration phase before the session. In this process, the software set the minimum and maximum values that correspond to the user's neutral face shape deformation, in other words, no face shape deformation. We need to capture this in the most natural way possible, so we ask the user to watch a picture of nature scenery during 15 seconds without acknowledging there is a calibration process working on the background. When a user watches the picture it is likely, that we can get the least face shape deformation, or in other words, the most neutral face of the user. If we show a video or tell them they are been analyzed is more likely to track noise in the face shape deformation.

Step 2. Content experience

The user will watch a video or play a game. During this step, the software tracks and saves both user's eye gaze and face shape deformation. When the software detects a face shape deformation, we add that time instance to the user's personal narrative timeline.

Step 3. User feedback

After the content experience is finished, we ask the user to describe a detailed story of the video or game. This step works as our narrative model so we can correlate the user's memory with the narrative timeline predicted by the software.

Step 4. Narrative timeline.

The user will be able to see a narrative timeline that the software predicted. The timeline contains each frame when the user experienced a change in face expression including the eye gaze position on the screen and his face shape deformation.

3.3 Narrative Timeline

The high definition face tracking available in Microsoft Kinect 2 provides up to 17-face-shape animation unit (AU's) values based on the user's face shape deformation. This animation values have a strong relation with the Facial Action Coding System published by Paul Ekman [8]. Ekman's coding system associates different face muscle deformations with action units. The sum of this action units leads to a code number that represents an emotion. For example, if we relate the Lip Corner Puller with the Cheek Puff we can presume the user is happy.

We create the narrative timeline using the following 17-face-shape animation units:

AU Index	Name
1	Jaw Open
2	Lip Pucker
3	Jaw Slide Right
4	Lip Stretcher Right
5	Lip Stretcher Left
6	Lip Corner Puller Left
7	Lip Corner Puller Right
8	Lip Corner Depressor Left
9	Lip Corner Depressor Right
10	Left Cheek Puff
11	Right Cheek Puff
12	Left Eye Closed
13	Right Eye Closed
14	Right Eyebrow Lowerer
15	Left Eyebrow Lowerer
16	Lower Lip Depressor Left
17	Lower Lip Depressor Right

Table 1. Animation Unit (AU) indexes

If the face shape is deformed, the software adds it to the narrative timeline. To tell if the face was deformed we calculate the difference between the user's neutral face shape deformations, in other words, no deformation, and compare it with the current face shape deformation for each of the seventeen indexes. The software considers a change in face shape when the index is out of range of the neutral deformation minimum and maximum values. The software uses a threshold value to know the percentage of changed indexes, and then determine if there was a face shape deformation. We can modify the threshold value during the data analysis for setting the best approximation possible to match with the user's story.

3.4 Face Shape Deformation Measurement

There are two different ways of displaying the face expression data. The first one is through a radar graph (Figure 1) that shows the seventeen animation units values for a current face shape deformation (magenta), and the minimum (green) and maximum (blue) neutral values. With this graph, we can easily know which animation units are out of the neutral range.

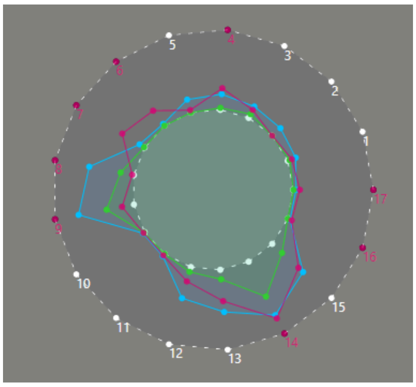

Figure 1. The figure shows how indexes such as 6 and 7 are out of the neutral range. The magenta colored line represents the current deformation while the green and blue represent the minimum and maximum neutral deformation values respectively.

Additionally, a second way to represent this data is transforming the AU's into a face graph, to display in a general way, the user's face expression for the current frame in the narrative timeline. This second way of data representation will help us associate specific face expressions with specific events. For example, for a certain user a smile could mean the trigger for a memorable event

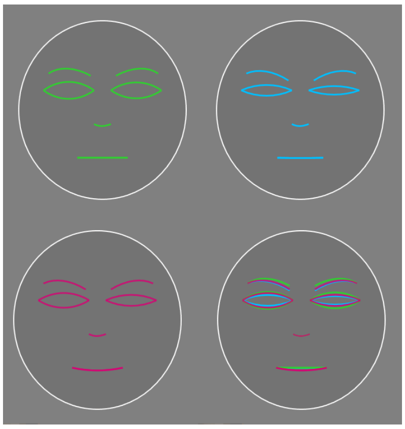

Figure 2. Face shape deformation as a face graph. The green and blue faces represent the minimum and maximum neutral face shape deformation respectively. The magenta one represents the deformation for the current frame. The last one is a superimposed image of all face graphs.

3.5 Eye Gaze Measurement

The way of measuring eye gaze is using the coordinates on the screen that the eye sensor provides. We use the position of the eye gaze to verify the user's story with a frame in the narrative timeline. For example, if the user recalls some object or character in his story, we can identify it with a mark in the narrative timeline frame, which indicates the exact position of the user's gaze at that moment.

4. Implementation

4.1 Preliminary Experiment

The purpose of this experiment was to ensure the Narrative Analyzer software worked properly. To prove this, four participants were subject to this preliminary experiment, which consisted in watching an animated shortcut movie. The threshold used for detecting the face shape deformation was the same for all participants.

Results came out satisfactory because the user's timeline, and the stories they wrote, correspond with the narrative timeline predicted by the software. As we mentioned before the narrative model consists in the last step of the workflow software where we ask the user to write the recalled story. This story, which is the user's narrative, should match with the predicted narrative timeline created by the software.

In general, there are different ways of classifying how the user recalls the story. The first is a whole description of what they have just watched. The second one is describing a partial sequence of the movie and finally, and most importantly, the description of a single event. Here is an example of the obtained results:

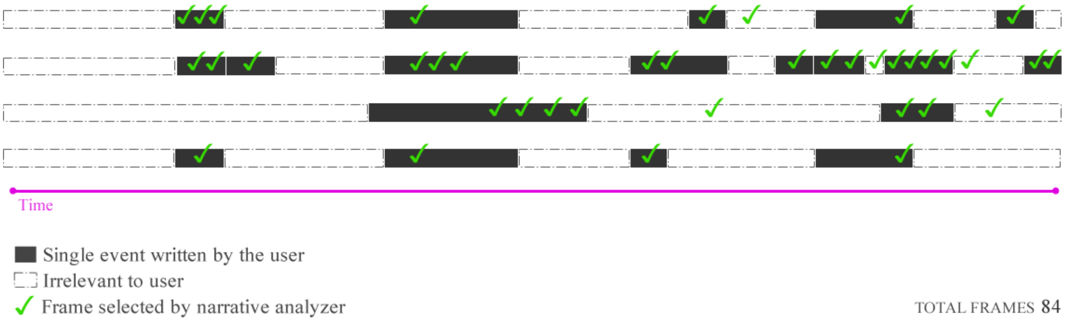

Figure 3. Correspondence between the predicted timeline generated by the software and the story remembered by the user

4.2 First Experiment

This is a second session using the Narrative Analyzer on a simple shortcut film. It is important to make emphasis on the correspondence between the user's personal story and the automatically generated narrative timeline.

Unlike with the previous experiment movie, this shortcut contains many partial sequence events, adding many frames to the timeline and in consequence make unclear to analyze the performance in Figure 4. Among all users, we have 28 single events. Single events are the most important because they exactly represent what the user remembers. In this case, the Narrative Analyzer recognized 23 of them. In addition, we can see that there is 167 successfully recognized frames, between single and movie partial description events, against seven false positives.

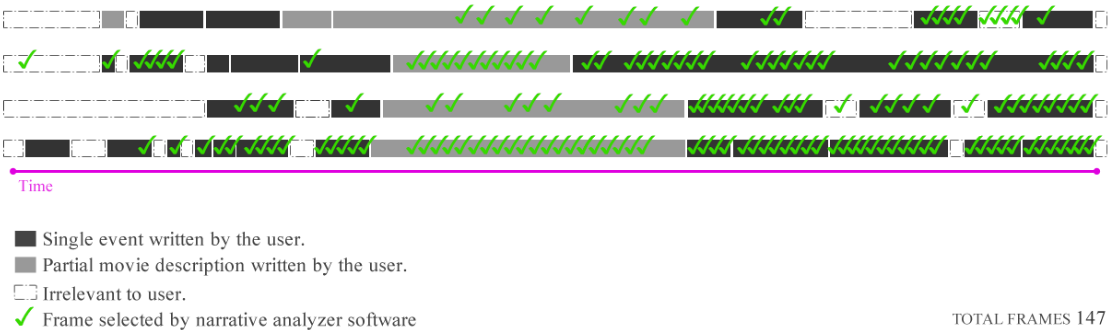

Figure 4. The graph represents performance of the Narrative Analyzer software. Single events are short in time but the software still could predict the majority of them. The dotted area represents parts of the movie that the user did not remember.

User	Threshold	Num. of frames generated	Frames correctly predicted vs false positives	Frames matching with user's story single events	Other frames matching with user's story events
1	3	19	15 / 4	3 / 5	1 / 3
2	4	45	44 / 1	4 / 6	1 / 1
3	3	38	36 / 2	5 / 5	1 / 1
4	3	72	72 / 0	11 / 12	1 / 1

Table 2. Results of the second experiment comparing the effectivity of the generated narrative timeline

4.3 Gameplay Experiment

The previous sessions used a movie to generate the user's narrative timeline. In this new experiment, we analyzed five participants. The purpose was to find the differences between watching a movie and playing a puzzle style multiplayer game in real time. While watching a movie generates changes in a user face expression, the real time feedback of playing a game causes additional reactions in the user. In the next figure, we compare the change rate per animation unit for the previous movie session and the new gameplay session. We can notice that a user's face expression has more changes while playing a game than simply watching a movie.

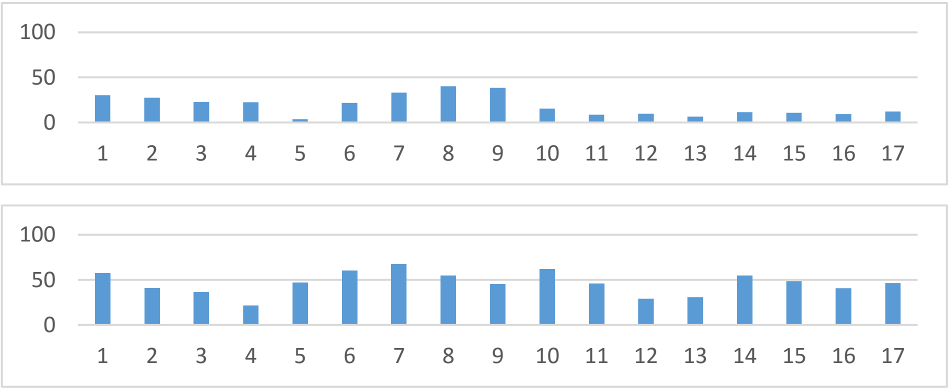

Figure 5. Change rate per AU, for a movie session and a gameplay session respectively

In this gameplay session, we add one more step in the narrative analyzer software's workflow. After the users have written their personal stories of the played game, we asked them to select all remembered frames showing a replay timeline. The goal of this additional step is to verify the story with what the user claims to remember. With the user-selected frames, we have the frames that the narrative analyzer software should select.

Firstly, we find out which are the indexes with more variation. For this, we compare the change rate for all the AUs analyzing all frames tracked by the software and all frames picked by the user.

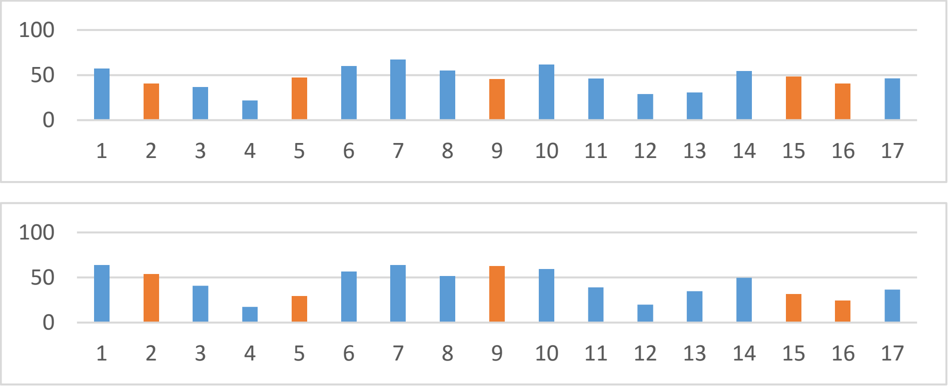

Figure 6. Change rate for all 17-animation-units. The upper graph considers all frames in the session; the lower graph considers the frames picked by the user.

From the previous figure, we can know that the AU's that vary the most are the Lip Pucker, Lip Stretcher Left, Lip Corner Depressor Right, Left Eyebrow Lowerer and the Lower Lip Depressor Left. We clearly notice that the mouth is the most relevant part of the face when referring to face shape deformation.

For each user we get the frames generated by the narrative timeline and intersect them with the frames selected by the user. The first graph column considers all the seventeen AU's indexes and the second graph column considers only the AU's mentioned in Figure 6. Unfortunately, the percentage of intersecting frames is low for all users. The reason is that when the user sees the replay of the game, his memory may change from the previous written events hence, the low amount of intersecting frames. Even the frames selected by the user are the ones that he claims to remember, this is not always the truth. The grand truth is not easy to find, and it is a complex problem to solve.

4.4 Second Gameplay Experiment

For this last experiment, we use a single-player platformer style game for analyzing the user. We tested the accuracy of the software with the following results:

User	Threshold	Num. of frames generated by the software vs total available frames	Num. of intersecting frames w/user selection	Frames correctly predicted vs false positives	Frames matching with user's story events
1	6	40 / 80	11 / 14	36 / 4	6 of 7
2	3	42 / 134	5 / 15	18 / 24	4 of 7
3	2	14 / 52	2 / 11	12 / 2	4 of 7
4	4	26 / 220	2 / 10	16 / 10	4 of 6
5	4	21 / 177	5 / 29	13 / 8	4 of 5

Table 3. Effectivity of the generated timeline in a real time gameplay session

Example of events written by users	Num. of users who wrote this event	Picked effectivity
The player received a reward	4 of 5	100% (4 of 4)
The player received damage	5 of 5	80% (4 of 5)
The player failed the stage	4 of 5	100% (4 of 4)

Table 4. Example of events written by the user and the effectivity of the software

To consider a frame as correctly predicted we matched each of the frames selected by the software against the user's written story, including the position of the eye gaze when the story had stated specific objects. We notice that 95 out of 143 frames matched with the users' story with only 34% of false positives. For the events that the user wrote, 10 out of 32 did not appear on the generated narrative timeline, meaning 31% of failure.

As opposed of our thinking during the previous game session experiment, for the intersecting frames between the ones selected by the user and the ones selected by the software, we can say that the former frames are not necessarily accurate when matching the story with the generated narrative timeline. The important thing to point out is the presence of the events that the user wrote in the generated timeline, in this case around 70% of success.

4.5 Gameplay Sessions Further Analysis

The next analysis shows that there is a difference in face shape deformations for each different game. This means that for each unique experience, the animation units that we have to consider for generating the narrative timeline also have to change. In Figure 7, we can see the average face shape for each of the gameplay sessions:

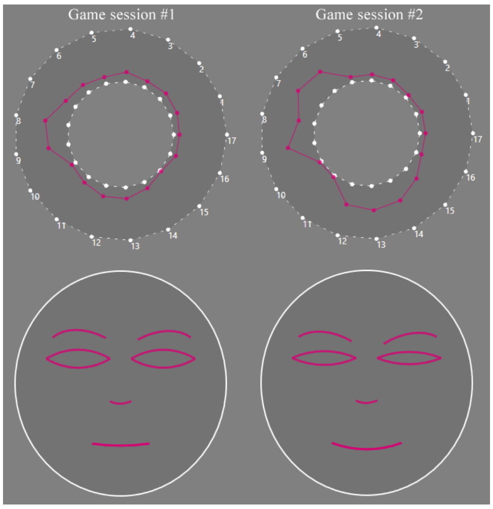

Figure 7. Average of all users face shape deformation. Left shows the face for a multiplayer puzzle style game, and right shows the face for a single platformer game.

With this in mind, we can see that the user's reaction for each game is different. For instance, in both of the game sessions the users' smile exists, but in the second game is clearly bigger. In Figure 8 we compare the percentage of average changes for each AU, we notice there is a similar tendency for the indexes with the exception of numbers 2, 4, 8, 9 and 10, which not unexpectedly, are mouth deformation's animation units.

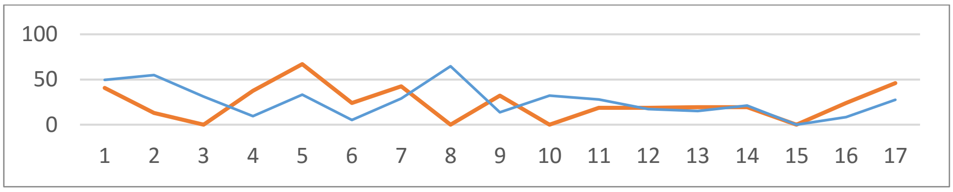

Figure 8. The blue line represents the game session #1; the orange line represents the game session #2

This analysis shows an interesting result as we could create a game classification based on face expressions. For example, the first face in Figure 7 represents the average face for a puzzle style multiplayer game; meanwhile the second face of the same figure represents the average face for a single player platformer game.

5. Other experiments

For improving the accuracy of the narrative timeline generated by the Narrative Analyzer software, we added a variant of the experiment by letting the users pick the frames they remembered just after finishing watching the video. Then we could compare the selected frames with the ones automatically selected by the software. The purpose was to find a pattern along this process. Even the users selected the frames they remembered; the selections may not be true memorable events. We implemented the following data learning analysis methods for these experiments using Accord Framework [9]. Although the results were not satisfying it is important to mention them.

5.1 Support Vector Machine

Given a group of data, this machine learning tool helps classifying and creating a model for grouping any new given data. The purpose of using the SVM analysis was to find a way of classifying face shape deformations and try to associate specific facial expressions when the user recognized a memorable event. For this, prior to feeding the data to the support vector machine, we try reducing the dimension of it by applying the Principal Component Analysis (PCA). In Figure 9, we can notice that the division of the data was not successful:

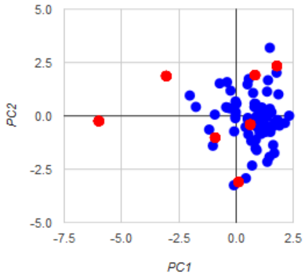

Figure 9. The red spots are the frames selected by the user, and the blue spots are all the others.

Due to the nature of the data, the SVM technique cannot work correctly, as it was not possible to find a good classification for the seventeen different indexes.

5.2 K-Means

With the same purpose of classifying data, the idea of using K-Means was to create clusters of the different faces recorded during the experiment session. If we use the means generated for the current session, we can find some patterns that correctly match the user's story and associate it with a face expression. The counterpart is that if we use the same means to analyze a different session, the results are random, making it impossible for improving the accuracy of the generated timeline. In the next figure, we used 80 clusters for the K-means calculation but still we cannot see any patterns in the frames the user claimed remembering:

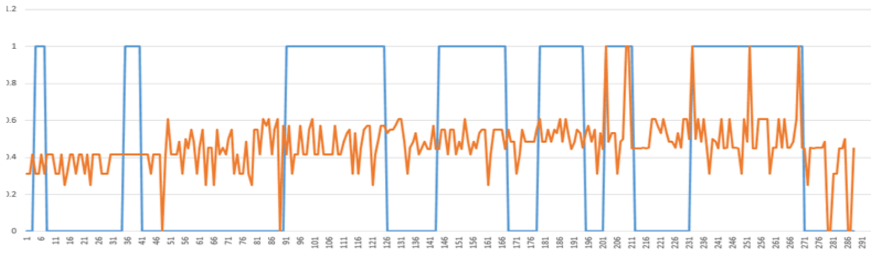

Figure 10. The blue peaks show the frames selected by the user (1 meaning remembered, 0 meaning not remembered). The orange graph represents the means of the face shape deformations among time.

5.3 Pupil Measurement

Using additional sensors could be a solution for improving the accuracy of the software. There is some research about how emotional stimuli affects the size of the eye pupil's diameter.

In particular, Timo and Veikko's research [10] demonstrates that the pupil's diameter suffers big changes in size when the user is subject to emotional stimulation. We tried measuring the changes in the eye's pupil diameter and its association with the face shape deformation. Unfortunately, there was no related pattern found between both measurements.

In Figure 11, the orange line shows the pupil's diameter size among time. The green graph shows the frames that the user selected as remembered instances. As we can see, there is no relation between the regions selected by the user and the pupil's diameter. Even discarding the noise data, the pupil's diameter keeps changing randomly all along the experiment session.

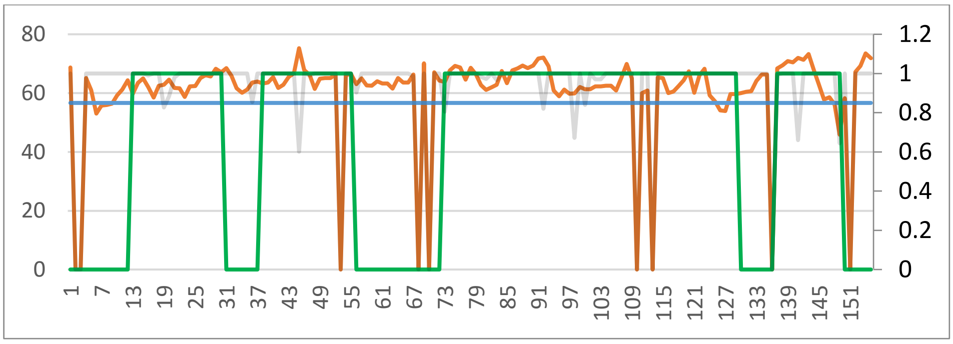

Figure 11. The green peaks show the frames selected by the user (1 meaning remembered, 0 meaning not remembered). The orange graph is the pupil diameter size among time.

6. Conclusion

The purpose of the Narrative Analyzer software is to predict the user's narrative through a narrative timeline generated automatically. We could produce a narrative timeline that matches close enough with what the user considered the most important events of the experienced content, based on the user's attention, using eye gaze data, and the user's reactions using face shape deformations. After executing all the experiments detailed in this research, we found out that the generated narrative timeline matches a major number of events remembered by the user.

For the case of using the software in simple content such as short movies or video clips, the narrative timeline that the software generates matches with most of the memorable instances for the user.

On the other hand, when we use the software in a gameplay session, results vary from game to game. The difficulty, type (single or multiplayer) and other factors of the game affect directly on the user's face expression, which makes it quite difficult to find a pattern for analyzing the data. For example, if the game is too complex, the number of false positives increase.

Another point to mention is that the story the user writes and the frames the user selects as remembered events do not necessarily match. The reason of these conflicted results is that the events that a user remembers are not easy to recognize even by the same user. To find the truly memorable events is a complex problem to solve.

Our tool is a good start for the difficult task of finding the grand truth about a person's personal narrative. Considering face shape deformations, or in other words, face expressions, combined with the eye gaze can make it easy to understand what really matters to the user and what we need to consider when trying extracting narrative events.

References

[1] J. Hillis Miller: 'Narrative', Critical Terms for Literary Study. Frank Lentricchia and Thomas Laughlin. Chicago, 1995: 66-79

[2] Aristotle, Poetics.

[3] Henry Jenkins, Game Design as Narrative Architecture, in First Person. New Media as Story, Performance, and Game (eds.) Pat Harrigan and Noah Wardrip-Fruin. MIT Press, Cambridge 2003 (in press).

[4] Cahill L., McGaugh J. L, Mechanisms of emotional arousal and lasting declarative memory, Trends Neuroscience, 1998, Volume 21, 294-299.

[5] Dere E., Pause B., Pietrowsky R, Emotion and episodic memory in neuropsychiatric disorders, Behav. Brain Res, 2010, 162-171.

[6] Katie Salen and Eric Zimmerman, Rules of Play, Game Design Fundamentals. MIT Press, 20

[7] Wei-Ting Peng, et al. A User Experience Model for Home Video Summarization Advances in Multimedia Modeling, Volume 5371, 2009, 484-495.

[8] Paul Ekman and Wallace Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, 1978

[9] C. Souze, Accord.NET Machine Learning Framework: Available at http://accord-framework.net/, version 3.02

[10] Timo Partala and Veikko Surakka, Pupil size variation as an indication of affective processing. Int. J. Human-Computer Studies 59, 2003