On the automatic preference ordering of digital photo’s through viewer behavior analysis

Simon Epskamp – s0088617
simon.splash-infba@gmail.com

ABSTRACT
In this paper we look at people looking at photos. The relation between how long a viewer watches a photo and how much the viewer likes that photo is researched. Likewise the relation between the facial expression that the viewer shows while watching and viewer preference is looked at.

An experiment where eight participants were shown two sets of photos is conducted. Participants look at the photos at their own pace, then create their preferred ordering per set.

A tool is used to classify facial expressions of viewers into emotions. Only the happiness emotion turns out to be detectable reliably.

An analysis of the acquired data shows that the time watched doesn’t indicate preference, but facial expression does.

Keywords
Photo classification; viewer preference; facial expression; automatic emotion detection; photos.

1. INTRODUCTION
With digital photography coming available to the general public the price of taking a picture has decreased to virtually zero. Once the equipment (camera, computer) is bought, taking another picture doesn’t come at an added cost, since it can be viewed over and over again before deciding if it should be printed.

This has led to a problem: overproduction. People often make lots and lots of photographs, without going through the hassle of sorting them out later. This leads to the pictures being rerouted to the big ‘pile’, to never be watched again.

However, the personal computer that made this situation possible might also be the key to it’s solution. It can easily be extended with ways to observe its users as they are watching their photographs for the first time. Analysis of the observed behaviors might then indicate which photo’s are liked, and be stored, or even discarded. The ideal is a system that, through user observation, would create an ordering in the photos that are designed for this specific purpose.

Of these behaviors, two that could easily be observed were selected for study. The first one is ‘relative time watched’. This behavior can easily be detected when the person viewing the photographs is also deciding when to show the next one. Simply recording the timestamps at which changes occur is enough. The second selected behavior is ‘showing emotion through facial expression’. Building our own facial expression analyzer isn’t necessary, since there are several solutions on the market that are designed for this specific purpose.

In this paper we will first more formally define the research questions. A look at prior research projects with similar subjects follows, which includes a more in-depth look at the techniques behind facial expression recognition. The methods for detecting the chosen variables, the experiment, and how the collected data is analyzed are next, followed by the conclusion and discussion of the results.

2. RESEARCH QUESTIONS
The main research question is:
“Do the relative time that a person watches a photo, and his/her facial expression during this time, relate to how much the viewer likes the photo?”

This question will be partially answered through researching the above mentioned behaviors, guided by the following sub-questions:
“Does an ordering of photographs by time watched relate to the ordering a viewer would choose him/herself?”
“Can viewer’s facial expressions be reliably be detected, and mapped to emotions?”
“Is it possible to relate detected emotions to viewer preference?”

3. PRIOR RESEARCH
In this section a sense of what is going on in the field is established. First, a review is given concerning the current state of the art in emotion detection, since this is one of the primary basis for automatic tagging of media based on viewer response. In the second subsection, a paper that describes research in this direction is discussed.

3.1 Automatic emotion detection
Automatic detection of emotion in a video stream is an active field of research. [3] [4] [5]

In order to achieve automatic facial expression recognition, it must first be determined whether or not a face is present in an image. So-called face-detection methods can be divided into two subgroups. [6] Firstly, knowledge-based methods try to implement human knowledge of what a face looks like. Encoding this knowledge is the challenge, it might be done as...
contrast distribution, or skin texture or color, or the shape of organs like the mouth. Secondly, template-matching is based on a default face pattern, that is matched to all regions of an image. This template can be defined by experts, or be learned from example images.

Once the location and rotation of faces in an image is known, the specific expression data must be extracted. [7] For this task, template-matching with more detailed templates can be used. In this case, a generic template is matched to the face as closely as possible. Alternatively, a feature-based method can be used that tries to find certain landmark points in the face. Example landmarks are the corners of the mouth or the start points of the eyebrows. Their relative positions make up the expression data in this case.

Lastly the extracted facial expression data must be mapped to a certain or several emotions. [7] For describing the movements of the face, the Facial Action Coding System [8] is the most widely used method. This system describes every possible facial expression in terms of 44 so-called Action Units. An example of an AU is the “Lower Lip Depressor” that indicates how much the lower lip is pulled down.

The emotions that are detected are usually the six “universal emotions” as described by Eckman: happiness, sadness, surprise, fear, anger, and disgust. [9] However, there is as of yet no universally accepted method of mapping the action units to the six basic emotions.

3.2 Multimedia indexing

Applying automatic behavior detection and analysis to classification or indexing of multimedia seems a rare idea. Relatively little prior research can be found on the topic. The one paper that does deal on this subject is the one by Kowalik et. al. [10]. In this paper a system is proposed that allows for automatic tagging of video sequences. Viewers are watched using a digital camera, and feature points are extracted. These are classified to recognize several action units that are related to the emotions of joy, sadness and anger. Sadness and anger is recognized with high correctness, but sadness not so much.

The process is then tested with several test subjects, that are asked to indicate funny parts in a video, which are then matched to the output of the joy analyzer. The authors conclude that funny parts can successfully be recognized since the occurrence of indicated funny parts fell in intensity peaks of the joy analyzer 90% of the time.

In the end this research feels incomplete. A lot of time is spent discussing the inner workings of the system, but in the end only the application of the joy detector is tested. No thought is given as to how the other detectors might perform, how the result might be combined, or how a specific part of a clip might be easier to retrieve using the data obtained.

4. METHODOLOGY

A user study will be conducted to attempt to answer the research questions. A software prototype was built to aid in this. In the next sections the chosen behaviors, and how they are detected are discussed in more detail, followed by the experiment’s details.

4.1 Detection of chosen behaviors

4.1.1 Showing emotion

Facial expression recognition is selected as a viable method of detecting emotion in the viewer. People might subconsciously use facial expressions such as smiling, or body posture like pointing when they watch their photographs [11]. These might be linked to liking/disliking a photo.

A facial expression analyzer at Amsterdam University called eMotion will be used. This system utilizes a knowledge-based facial feature detector for face detection, and template matching for extraction of the expression data. This data is in the form of Motion Units that are similar but not equivalent to the Action Units described before, since they also include motion, the direction and speed in which a certain Action Unit is moving. These MUs are then mapped to emotions using an automatic classifier, that is trained on a large set of MUs with known corresponding emotions [12] [13]. The system rates expressions on the six basic emotions, each on a scale of 0-100 for each frame. A seventh ‘emotion’: neutral is introduced to represent frames on which no other emotion is detected.

What remains to be decided is how these emotions will be mapped to viewer preference. Kowalik et. al [10] selected happiness, sadness and fear as interesting emotions for their video tagging system, but don’t attempt to combine these emotions in one ‘interest’ rating. A feasible assumption however is that the ‘positive’ emotions happiness and surprise indicate a liking of the photo, while the ‘negative’ ones, sadness, fear, anger and disgust indicate a dislike.

4.1.2 Relative time watched

How long people look at a certain photo might be a clear indication of how much they like it. The assumption here is, of course, that people will watch a photo they like more for a longer time.

Simply saving the time span a photo was shown on screen is enough to track how long a photograph is watched.

In real-life situations this would have to be combined with the face detection and whether the computer program showing the photo had focus, to make sure someone is watching. In our simple test-setup only saving the time will be enough, since participants will always be looking at the screen, interacting with the program.

4.2 Experiment setup

Eight people participated in the experiment. Of these, six were male and two were female. All are aged between nineteen and twenty-five.

Participants were seated behind a computer, and given a small introduction. They were only told that the research would be about digital photographs. Nothing more about the study was explained, as to influence their behavior as little as possible. Next the practical details of the study were explained to them.

These consisted of the fact that they were presented two sets of 16 and 18 photographs each. The participants chose themselves when they wanted to see the next photo (see Figure 1, the button labeled ‘volgende’ selects the next photo). Afterwards, they were asked to select which photographs they would ‘keep’ if it were their own photos, and which ones they would throw out. (see Figure 2) As a general guideline, they were asked to throw out about half of them. Furthermore, they were asked to order the photographs from ‘best liked’ to ‘least liked’, where it was impossible to move photo’s indicated in the previous phase as ‘throw out’ above the ‘keep’ photos. (see Figure 3)
All this was implemented as a Java program. The participants were then shown a small demo with three photos per set to make the process more clear. The real process was then started. The participants were told to relax and take their time, then left alone in the room, to induce a more natural viewing state.

During the photo viewing stage, the participants were recorded using a webcam (Figure 4), so that their facial expression could be analyzed later. The timestamps for starting/stopping to watch all photographs, which ones they wanted to keep/throw out and their final order was also recorded.

When the participants were ready, they were shown their final order, and asked to explain, in their own words, why they liked or disliked certain photographs.

4.2.1 Photo sets.
Two sets of photos were selected for this study. The first set consisted of a subset of the “International Affective Picture System” (IAPS). This is a set of photos that was rated by a wide selection of subjects on two primary scales:

“one of affective valence (ranging from pleasant to unpleasant) and one of arousal (ranging from calm to excited)” [14]

The reason this set was used is that it provides a wide selection of photos that none of the participants have seen before, that span the complete spectrum of induced emotions, and that are rated on these emotions so that a proper distribution in the set can easily be achieved. Therefore, 16 photos were selected from the following categories:

- Medium valence and low excitement (boring photos, i.e. a spoon)
- Medium valence and high excitement (exciting photos, i.e. lightning)
- High valence and low excitement (cute photos, i.e. puppies)
- High valence and high excitement (pleasant, exiting photos, i.e. skydivers seen from the perspective of one)

Note that no photos rated with low valence were selected. The reason is that this category contains photos that are so shocking that the fear existed that they would have influenced further viewing behavior.

The reason 16 and 18 photos per set were selected is to find a balance between having as much data as possible, yet preventing that participants became bored and would change their behavior accordingly.

The second set consisted of photographs selected from a personal collection, of a single event. This set was chosen because the photos in this set are closer to the kind of photos the system would ultimately be used for.

Again, the range of induced emotions was to be as wide as possible, therefore the selected photos ranged from very ‘interesting’ (i.e. multiple people smiling at the camera, doing something interesting), to very ‘boring’ (i.e. out of focus photographs, boring subject such as way signs). This set contained 18 photographs.

5. DATA ANALYSIS
Besides facial expression and relative time watched the participants provided us with their choice whether they would like to keep or trash all photographs, and the order that has the most like photo on top, and the least liked photo on the bottom. This ordering will hereafter be referred to as the chosen ordering. This chosen ordering has a certain barrier up to which the photos were marked as ‘keep’, and under which the photos were marked as ‘trash’. This barrier will be referred to as the trashbarrier from now on.

The fact that we have the chosen ordering allows us position this barrier at different places. For example, we might place it halfway from the top, to simulate an equal number of photos in the trash and keep categories. This barrier will be referred to as the halfway barrier from now on.

5.1 Relative time span watched
Since the exact time at which a participant started watching and stopped watching a certain photograph was recorded, the data consists of the number of seconds each photograph was watched.

The photos can then be sorted by the number of seconds watched, from most to least, to form a preference ordering. The
question that remains is how to compare this ordering to the chosen ordering. This is done by dividing both orderings at either the trash barrier or the halfway barrier, thus creating the categories of ‘keep’ and ‘trash’. It is then computed what percentage of the photos in the ‘keep’ section of the chosen ordering is present in the time ordering, we call this number %keep. The same is done for the ‘trash’ section, producing %trash. These numbers are then averaged to produce the final percentage of photos that was classified correctly: %total = (%keep + %trash) / 2.

If the time watched doesn’t correlate to viewer preference at all, %total will be about 50%. If the assumption that longer viewing times indicate a higher preference is correct, %total will be higher. Numbers lower than 50% indicate the relation is the other way around, shorter viewing times indicate more preference.

5.2 Facial expression

The eMotion software used (Figure 5) rates facial expression on six emotions: happy, surprised, angry, disgusted, fearful and sad, each on a scale from 0 to 100. However, when using the software it quickly became apparent, that, unfortunately, the only emotion that was rated reliably was happiness. The other emotions all were often detected when no expression was observed.

This method is called the limit method. The limit is necessary to weed out flukes caused by the eMotion software. These flukes are caused by small fluctuations in happiness detected in the video stream (probably caused by noise or grain). A value of 0.16 for the limit turned out to weed out most so-called ‘false positives’.

The photos that remain after this selection can be analyzed in the same way as before.

5.3 Combination

If both methods give a positive result, the combined results can then also be checked to see if this improves the final ordering. The variables will first have to be normalized, therefore the photo that induced the most positive emotion will be rated 1 on this scale, the most negative one 0, with the others in between based on a linear scale. The viewing time will be normalized in the same way.

The addition of these variables then leads to a third ordering, which can be divided and analyzed.

6. RESULTS

6.1 Relative time watched

When ordering the photos by relative time watched, and utilizing the halfway barrier to divide the chosen ordering and time ordering into trash and keep, the number of photographs classified correctly for all participants was 62 out of 120 photos for the keep category, and 73 out of 128 for the trash category. This amounted to a total percentage of 56.8% classified correctly. (see Table 2).

Table 1 - relative time watched - trash barrier

<table>
<thead>
<tr>
<th>kept class. correctly</th>
<th>total kept</th>
<th>trash class. correctly</th>
<th>total trashed</th>
<th>%total</th>
</tr>
</thead>
<tbody>
<tr>
<td>83</td>
<td>130</td>
<td>70</td>
<td>118</td>
<td>56.79%</td>
</tr>
</tbody>
</table>

When utilizing the trash barrier 83 out of 130 photos for the keep category were classified correctly, and 70 out of 118 for the trash category. This amounted to a total percentage of 54.3% classified correctly. (see Error! Reference source not found.)

Table 2 - relative time watched - halfway barrier

<table>
<thead>
<tr>
<th>kept class. correctly</th>
<th>total kept</th>
<th>trash class. correctly</th>
<th>total trashed</th>
<th>%total</th>
</tr>
</thead>
<tbody>
<tr>
<td>62</td>
<td>120</td>
<td>73</td>
<td>128</td>
<td>54.34%</td>
</tr>
</tbody>
</table>

6.2 Facial expression

6.2.1 Ordering

When utilizing the trash barrier 58.05% of photos was classified correctly (Error! Reference source not found.). When utilizing the halfway barrier 52.47% of photos was classified correctly (Table 4).

Table 3 - happiness ordering - trash barrier

<table>
<thead>
<tr>
<th>kept class. correctly</th>
<th>total kept</th>
<th>trash class. correctly</th>
<th>total trashed</th>
<th>%total</th>
</tr>
</thead>
</table>
6.2.2 Limit

In this analysis, only photos that have a happiness value of above 0.16 are considered, as explained in section 5.2.

When comparing the photographs that meet this limit to the keep category, using the halfway barrier, we see that 25 photographs meet this limit, of these 19 are present in the keep category. A 76.9% success rate.

Furthermore, when observing the results, it becomes clear that some photos that were shown after a photo with a high happiness rating also have a high happiness rating. See Figure 6 for an example. This image shows a piece of a happiness ordering, with the bars representing the amount of happiness detected while showing that image, and the number representing the position of this image in the chosen ordering. (0 being at the top) It becomes clear that the bottom picture has an above-average happiness rating considering its chosen position. This phenomenon can be observed in multiple places throughout the data-set. And explanation might be that the happiness rating as shown here is averaged over the exact time the photo was onscreen. However: a participant will need some time to react to a photograph as it appears onscreen. While still smiling, he/she will select the next photograph. (afterglow)

![Figure 6 - portion of a natural ordering](image)

To compensate for the reaction time and afterglow, the window for analyzing happiness for a certain picture was shifted to start when the picture was onscreen for 1 second, and end 1 second after it disappeared. This adjustment causes the above classification to rise to 22 out of 26: 84.6%. The data-set is too small to see whether this is a genuine effect, or a coincidence.

7. CONCLUSION

Considering the time watched variable, we can safely say that in this study this variable does not relate to how much a participant likes a certain photograph. We draw this conclusion because the percentage of correctly classified images is close to the 50% a random ordering would achieve.

Considering the happiness rating, this didn’t turn out to be valuable for creating a complete ordering, since participants simply don’t emulate a lot of the time. This might have been caused by the eMotion software failing to recognize expressions at times. However, on manual inspection of recorded videos and the eMotion output for happiness it became clear that the software does consistently detect smiles. Having excluded a technical cause, the question remains to be answered if participants emulate more when shown different photosets, for example consisting of their own photos, or containing more vibrant subjects.

If participants do show happiness however, this is a strong indication that they will want to keep the photo, this was predicted correctly in 76% of cases, or even 84% when a reaction-time + afterglow compensation is utilized. Concluding we can state that we found that relative time watched is not a useful tool in determining viewer preference. Facial expression recognition however, shows promise.

7.1 Research questions revisited

When looking back on the research questions we can see whether or not we found an answer to them:

“Does an ordering of photographs by time watched relate to the ordering a viewer would choose him/herself?”

No, it does not. The ordering based on time watched doesn’t predict the chosen order any better than a random one.

“Can viewer’s facial expressions be reliably be detected, and mapped to emotions?”

Partially. When using the eMotion tool only the happiness emotion and corresponding expressions can reliably be detected.

“Do the detected emotions relate to viewer preference?”

Yes they do, a strong happiness rating indicates that the viewer likes that particular photograph.

Now that we’ve answered these sub questions, the main research question can be answered:

“Do the relative time that a person watches a photo, and his/her facial expression during this time, relate to how much the viewer likes the photo?”

The relative time watched does not relate to how much the viewer likes that photo in the end. The viewer’s facial expression does give an indication towards how much the photo is liked.

7.2 Further research

In this paper, a first step was taken towards automatically grading viewer preference for photographs. Further research might focus on fine-tuning the results obtained by detecting facial expression, by trying to obtain even better results from classifying the ‘happiness’ emotion, or taking into account other emotions too.

Furthermore, other factors like might be researched, such as leaning into the screen, talking about a photograph and showing emotion by body language. More factors could be discovered by means of a user study.

8. DISCUSSION

If this project were a bigger one, we would begin with a user study, by simply asking various people about their photo’s: which ones they like, and what they like about them. Observation might then indicate what behaviors they show while looking at, and talking about pictures they like or dislike, and which of these might be viable for automatic detection. Since the scope of this project is however quite limited, this had to be done on a hunch.

The eMotion software used didn’t turn out to as capable as hoped. The software showed heavy fluctuations in the angry,
disgust and sad departments, even when the participants were clearly showing a neutral facial expression. Luckily, the software did consistently detect happiness whenever the participant were smiling.

The reason why this happened remains to be guessed at. Valenti et. al, the developers of the software, don’t give any specific detection rates for the different emotions in the accompanying paper [15], so it is impossible to check if our findings are consistent with theirs. However, since global detection rates are high, and a defect of this kind would have been mentioned, it can be reasonably assumed that it is not consistent. Therefore, the reason might to be sought in one of these directions:

- The software was originally only tested with clips that were about 3 seconds maximum. Our clips were much longer, up to 5 minutes. The fitted face template sometimes shifted out of place over time, leading to worse detection.
- Image quality. Unfortunately we had to work with a rather old webcam that provided images of only 320x240 resolution, with a fair bit of grain. A better webcam might have improved detection rates.
- Face direction. as can be seen in Figure 4, participants were filmed slightly from above. Since they weren’t looking directly at the camera, this might have made realizable detection harder.

9. ACKNOWLEDGEMENTS
I would like to thank my parents for giving me the original idea for this paper. Furthermore I would like to thank my supervisor, Herman Koppelman for his constructive feedback.

10. REFERENCES


