Using facial recognition services as implicit feedback for recommenders

Toon De Pessemier
imec - WAVES - Ghent University
toon.depessemier@ugent.be

Ine Coppens
imec - WAVES - Ghent University
ine.coppens@ugent.be

Luc Martens
imec - WAVES - Ghent University
luc1.martens@ugent.be

ABSTRACT

User authentication and feedback gathering are crucial aspects for recommender systems. The most common implementations, a username / password login and star rating systems, require user interaction and a cognitive effort from the user. As a result, users opt to save their password in the interface and optional feedback with a star rating system is often skipped, especially for applications such as video watching in a home environment. In this article, we propose an alternative method for user authentication based on facial recognition and an automatic feedback gathering method by detecting various face characteristics. Using facial recognition with a camera in a tablet, smartphone, or smart TV, the persons in front of the screen can be identified in order to link video watching sessions to their user profile. During video watching, implicit feedback is automatically gathered through emotion recognition, attention measurements, and behavior analysis. An emotion fingerprint, which is defined as a unique spectrum of expected emotions for a video scene, is compared to the recognized emotions in order to estimate the experience of a user while watching. An evaluation with a test panel showed that happiness can be most accurately detected and the recognized emotions are correlated with the user's star rating.

CCS CONCEPTS

• **Information systems → Information systems applications**; *Data analytics*; Data mining.

KEYWORDS

Feedback, Emotion recognition, Facial analysis, Recommendation

1 INTRODUCTION

Many video services generate personal recommendations for their customers to assist them in the content selection process that becomes more difficult by the abundance of available content. In the application domain of video watching, the content is often consumed simultaneously by multiple people (e.g., a family watching together) or the device is shared by multiple people (e.g., a tablet is used by multiple people of the family). Moreover, in the context of a household, people may join and leave the watching activity while the video is playing. However, classic recommender systems are not adjusted to this dynamic situation. Typically, recommendations are generated based on the profile of the individual who initiates the video session. Family profiles can be created, but do not take into account who is actually in front of the screen or changes in the number of spectators during the video watching. However, manually logging in each individual user, one by one, would be time consuming and user-unfriendly. The same issues are applicable for the feedback process. Explicit feedback is not requested separately for each individual. For implicit feedback, such as viewing time, it is unclear to whom this refers. Moreover, since star rating systems are often ignored by the user, an automatic implicit feedback system would be more suitable.

This article presents a more user-friendly and practical approach based on facial recognition to log in automatically every viewer and fetch their preferences to compose a dynamic group for group recommendations. These preferences are derived from their implicit feedback, which is gathered automatically by detecting various facial characteristics during past video watching sessions. We evaluated this implicit feedback gathering by using facial recognition services based on a dataset of photos as well as with a user test.

2 RELATED WORK

Face detection is the technique that locates the face of a person in a photo. It is the prerequisite of all facial analysis and different approaches for the detection have been studied [4]. Facial recognition is the process of matching a detected face to a person who was previously detected by the system. In the study of Yang et al., this is also called face authentication and defined as the identification of an individual in a photo [18]. Related to this is the analysis of faces for the purpose of age detection and gender detection. Automatically detecting the gender and age group of the user (child, young-adult, adult, or senior) can be useful for initial profiling of the user. In this paper, various commercial services for gender and age detection are used: Microsoft's Facial Recognition Software: Face [3], Face++ [8], and Kairos [11]. Even more recognition services exist, such as FaceReader [15], but some are rather expensive or are not available as a web service that can be queried from a mobile device. So, the first research question of this study is: *"How accurately are these commercial services for age detection and gender detection in view of an initial user profile for video watching?"*

While watching video content or using an app or service in general, facial expressions of users might reveal their feelings about the content or their usage. In the field of psychology, the relationship between distinctive patterns of the facial muscles and particular emotions have been demonstrated to be universal across different cultures [6]. The psychologists conducted experiments in which they showed still photographs of faces to people from different cultures in order to determine whether the same facial behavior would be judged as the same emotion, regardless of the observers'

.

culture. These studies demonstrated the recognizability of emotions (happiness, sadness, anger, fear, surprise, disgust, interest).

Based on these concepts, facial expression recognition is described as the identification of the emotions. The automatic recognition of facial expressions, and especially emotions, enables the automatic exploitation of emotions for profiling and recommendation purposes. Therefore, the same three commercial services are used for facial expression recognition during video watching in this study.

Various researchers have investigated the role of emotions in recommender systems. Emotions can be used to improve the quality of recommender systems in three different stages [17]:

(1) The entry stage: when a user starts to use a content delivery system with or without recommendations, the user is in an affective state, the entry mood. The user's decision making process is influenced by this entry mood. A recommender can adapt the list of recommended items to the user's entry mood by considering this as contextual information [1].
(2) The consumption stage: after the user starts to consume content, the user experiences affective responses that are induced by the content [17]. Moreover, by automatic emotion detection from facial expressions, an affective profile of movie scenes can be constructed. Such an item profile structure labels changes of users' emotions through time, relative to the video timestamp [10].
(3) The exit stage: after the user has finished with the content consumption, the user is in the exit mood. The exit mood will influence the user's next decisions. In case that the user continues to use the content delivery system, the exit mood for the content just consumed is the entry mood for the next content to be consumed [17].

In this paper the focus is on the consumption stage. Users watch movies and their facial expressions are captured as a vector of emotions that change over time. The facial expressions, such as emotions, are used as an indicator of the user's satisfaction with the content. The assumption is that users appreciate a video if they sympathize with the video and express their emotions in accordance with the expected emotions.

Therefore, the second research question of this study is: *"Can facial expression recognition during video watching be used as an unobtrusive (implicit) feedback collection technique?"*

3 METHOD

To facilitate human-computer interaction for video watching services, an Android application has been developed with the following three subsequent phases: 1) User authentication with an automated login procedure and user profiling (gender and age) based on facial recognition to identify all people who are in front of the screen. 2) Personalized recommendations (group recommendations in case multiple people are in front of the screen). 3) Automatic feedback gathering while the chosen video is playing. Using the front-facing camera of the tablet/smartphone or a camera connected to a smart TV, the app takes photos of all people in front of the screen and sends requests to different facial recognition services.

Figure 1 shows the data flow. The research focus of this article is on the first and the third phase. In the first phase, the goal is to identify and recognize each face in the photo. For new faces, age and gender will be detected to create an initial user profile. In the third phase, the photos will be used for emotion recognition, attention measurements, and behavior analysis in view of deriving automatic feedback. The second phase, offering personalized recommendations, is used to help users in the content selection process and demonstrate the added value of facial expression recognition.

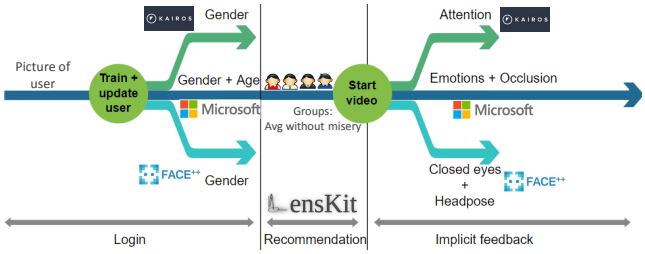

Figure 1: Data flow-3 phases: login, recommender, feedback.

3.1 Phase 1: User authentication and profiling

Although facial recognition is often used to unlock smartphones automatically, the applications in a group context, to identify multiple people simultaneously, are less common. In other words, facial recognition is used to give an answer to the question: "who is in front of the screen?". In a real world scenario, it can be several people, all of whom will be individually identified.

For the authentication of recurring users (who have been identified by our app in a previous session), our Android app uses Face Storage of the Microsoft service. This saves persons with their faces in a Person Group, which is trained based on the photos of the camera. This enables to link the user in front of the screen with one of the existing user profiles. For new users, the age and gender is estimated (Section 4.1). To cope with the cold-start problem, initial recommendations are based on these demographics.

In practice, user authentication and profiling works as follows. Using the app, users can log in by ensuring that their face is visible for the front-facing camera when they push the start button. A photo is made that is used as input for the facial recognition services. Recurring users are logged in automatically; their existing profile (age, gender, and watching history) is retrieved, and the new photo is a new training sample for Face Storage. For new users, a profile is created based on their estimated age and gender. After every login, the age estimation is adjusted based on the new photo. This update can correct age estimations based on previous photos, but also takes into account the aging of users when using the system for multiple years. This is especially useful for children who can get access to more content as they fulfill the minimum age requirements over time. Moreover, storing a photo for every session has the advantage that changes to the user's appearance (e.g., different hairstyle) can be taken into account.

3.2 Phase 2: Group recommendations

Group recommendations are generated by aggregating individual user models (consisting of age, gender, ratings, and watching history), one for every user in front of the screen. From the eleven

strategies proposed by Judith Masthoff [14], the "Average without misery" strategy was adopted in our group recommender algorithm. This strategy takes into account the (predicted) rating score of every user by calculating a group average, while avoiding misery by eliminating videos that are really hated by some group members, and therefore considered as unacceptable for the group. The Lenskit [7] recommendation framework was used to calculate these rating prediction scores and transform them into a Top-N recommendation list.

Besides personal preferences, other criteria, such as the age and historical viewing activities of the users, are taken into account. Age is modeled as classes of age ranges, firstly to filter out inappropriate content for minors, secondly for estimating the ratings for cold-start users based on other users of the same class. We used the age ranges that are also used by IMDb: <18, 18-29, 30-44, 45+.

The age of the users is used to determine whether a video is suitable for the viewers. For every video, the advised minimum age is retrieved from the Common Sense Media website [16]. If at least one of the users is younger than this age threshold, the video is marked as unsuitable for the group according to the average without misery strategy. Likewise, if at least one of the users has already seen the video, it is considered as unsuitable for the group since this person probably does not want to see the video again.

If a new user is present in front of the screen, i.e. a cold-start user, user preferences for a movie are estimated based on demographics. An estimation of the user's age and gender, as provided by the facial recognition services, is used to find users with similar demographics. The preferences of that demographic group (age & gender class) are used to estimate the preferences of the cold-start user. In case of an explicit rating for example, we use the mean rating of that demographic group for the movie, as mentioned by IMDb [9]. The mean rating provided by the demographic group is compared with the mean rating over all users for this specific movie. This difference (demographic group mean - global mean) indicates if the movie is less or more suitable for a specific age/gender.

3.3 Phase 3: Automatic feedback

Commercial services that perform emotion recognition, attention measurements, and behavior analysis are often based on the analysis of photos. Therefore, our Android app continuously takes photos from the users with the front-facing camera during video watching. Every second, a photo is taken and sent to the Microsoft recognition service for face detection and authentication in order to check if all viewers are still in front of the screen. Subsequently, for each identified face, the area of the photo containing the face is selected and the photo is cropped so that only one person's face is visible. Next, the cropped photo is sent to each of the three recognition services. Since photos are sent for every identified face, facial expressions will be recognized for all identified individuals in front of the screen.

For recognizing the emotions on the users' face, the Microsoft service was used. But these recognized emotions cannot be directly used as implicit feedback [2], since different videos evoke different emotions. One can assume that users appreciate a video if they sympathize with the video and express their emotions in accordance with the expected emotions. E.g., during a comedy scene users may laugh ('happy' emotion), whereas during a horror scene 'fear' can be expected. Recognized emotions that are not expected, might be due to external influences (e.g., other people in the room) or reflect contempt for the video (e.g., laughing with terrifying scenes of a horror movie). Therefore, unexpected emotions are not taken into account.

Thus, the similarity between the expressed emotions (=recognized emotions) and the expected emotions is calculated to determine the user's experience while watching the video. The expected emotions are based on the emotion fingerprint, which is defined as a unique spectrum of expected emotions for a video scene. For every second of the video, the emotion spectrum of the fingerprint specifies the probability value of each of the six possible emotions: anger, disgust, fear, happiness, sadness, and surprise. These emotion dimensions have been identified in the field of psychology [6]. So, the emotion fingerprint shows which emotions the video typically provokes among viewers at every second of the video. The emotion fingerprint is composed by aggregating emotions expressed by many users during watching this specific video. Section 4.4 explains in detail how the fingerprint of a video scene is computed based on an example.

The distance between expressed and expected emotions is calculated based on the euclidean distance between the values of these two emotion spectra for every second i of the video and each emotion j. For the expressed emotions, the output of the Microsoft service is used in our online experiment (Section 4.4) because of the results of the offline evaluation (Section 4.3). The similarity between expressed and expected emotions is calculated based on the inverse of the emotion distance and an additional constant to avoid a division by zero.

$$emotionDistance = \sqrt{\sum_{i=0}^{n}\sum_{j=1}^{6}(expected_{i,j} - expressed_{i,j})^2} \quad (1)$$

$$emotionSimilarity = \frac{1}{1 + emotionDistance} \quad (2)$$

Besides emotions, also the attention level and user behavior are analyzed during video watching as an additional implicit feedback mechanism. The Microsoft service has an additional interesting feature that recognizes occluded areas of the face. This *occlusion* is used to recognize negative feedback during video watching in case users respond to the video by holding their hands in front of their mouth or eyes (typical for shocking content).

Face++ is the only service that can detect *closed eyes*, which can be an indication of sleeping. Also the user's *head pose* is derived from Face++. Although other services can recognize the head pose as well, the estimation of Face++ showed to be the most accurate one. In case users do not want to see a scene (negative feedback), they might close their eyes or turn their head.

The Kairos recognition service offers a feature that represents the *attention* level of the user, which is estimated based on eye tracking and head pose. In our application, these behavioral aspects are combined into the *overallAttention* level by aggregating the service results over all photos taken during video watching. The overall attention level is calculated as the percentage of photos in which the user pays attention and following conditions are met: Kairos' attention level > 0.5, both eyes open, no occlusion, and

head pose angles are between the margins: 30 degrees for the yaw angle and 20 degrees for the pitch angle. The assumption is that the user is not paying attention to the video if one of these conditions is not met.

$$\text{overallAttention} = \frac{\#\text{Photos(attention \& eyes \& noOcclusion \& headPose)}}{\#\text{Photos}} \tag{3}$$

An *implicitFeedbackScore* on a scale ranging from 0 to 10 is calculated by aggregating the different facial analysis features. The similarity with the expected emotions has a contribution of six points out of ten points. The overall attention level counts for the remaining four points.

$$\text{implicitFeedbackScore} = 6 \cdot \text{emotionSimilarity} + 4 \cdot \text{overallAttention} \tag{4}$$

4 EVALUATION

Evaluations of commercial facial recognition services have been performed in literature, but are typically based on datasets with high-quality photos that enable an accurate recognition: sufficiently illuminated, no shadow or reflections, high resolution, and a perfect position of the face in the middle of the photo [2, 5]. In contrast, for facial recognition and analysis during (mobile) video watching, the front-facing camera of the device is used without flash, which yields not always ideal photos.

Therefore, we evaluated the three facial recognition services in an offline test (based on a publicly available dataset of photos) in Section 4.3 as well as in an online setting (with real users). For the evaluation of the age & gender estimation (Section 4.1), 46 users with age ranging from 0 to 66 were involved in our test. For the evaluation of the attention level (Section 4.2), we used 76 photos of our test users with different levels of attention. The evaluation of emotion recognition during video playback (Section 4.4) requires more time from the user and was therefore performed by only 20 users. Since the focus of this study was on age & gender estimation and emotion recognition, the group recommendations were not evaluated, and all users used the app alone.

The overall aim of this study is to improve the user friendliness of devices for video watching in the living room. This evaluation is the first step to reach the future goal of multi-user recognition and is therefore carried out with a tablet, in a rather controlled environment, with one person at a time. During the test, photos of the test user were taken with the front-facing camera of the tablet (Samsung Galaxy tab A6). If the tablet would have captured two people in the photo, the recognition process would be performed for both recognized faces.

To have a realistic camera angle, the users were asked to hold the tablet in front of them, as they would usually do for watching a video. The users were sitting on a chair and the room was sufficiently illuminated. However, no guidelines were provided regarding their behavior, head position, or attention; e.g., nothing was said about looking away or closing eyes. The photos taken with the front facing camera are used as input for the recognition services.

Table 1: Evaluation of gender & age estimation

	Kairos	Microsoft	Face++	Aggregation
Detection failed	4	2	2	2
Avg abs. age error	8.88	4.31	13.14	7.91
Median age error	6.0	2.9	11.0	8.1
Gender error (%)	11.9	15.9	13.6	11.3

4.1 Age & gender estimation

Firstly, the authentication was evaluated: recognizing the user who used the app in the past. The automatic authentication of the 46 users (login process) showed to be very accurate: 4 undetected faces with Kairos (9%), 2 with Microsoft and Face++ (4%).

Subsequently, for the recognized faces, the services were used to estimate the users' age and gender based on photos of the test users taken while holding the tablet. The estimated age and gender, as provided by the recognition services, were compared to the people's real age and gender. Figure 2 shows the differences between estimation and real age, sorted according to the real age of the users. The largest errors were obtained for estimating the age of children. Kairos and Face++ typically estimate the children to be older than they are. Table 1 reports the number of photos for which a detection was not possible, the average absolute age error, the median age error, and the percentage of photos for which the gender estimation was wrong.

The three facial recognition services were compared with a hybrid solution that aggregates the results of the three. For the age estimation, the aggregation is the average of the results of the three. For the gender, the three gender estimations are aggregated using a voting process. For each photo, the gender with the most votes is the result of the aggregation. This voting aggregation showed to be more reliable than each individual service for estimating gender.

Microsoft has the best results in the age test, so we decided to use only this service to estimate the user's age. For the gender estimation, the aggregation method is used.

So as an answer to the first research question, we can say that the facial recognition services provide an accurate age and gender detection for creating an initial profile for a cold-start user.

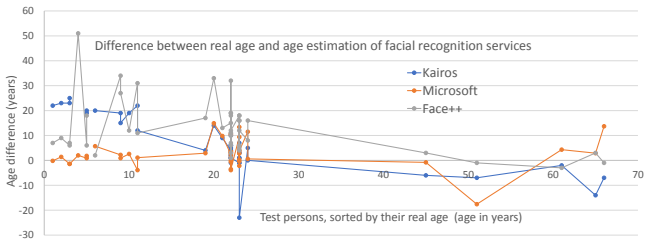

Figure 2: Age estimation using facial recognition services.

4.2 Attention level offline

The features that constitute the attention score of the user (equation 3) are evaluated based on a dataset that we created with photos of the users taken during the test. In addition, some photos were

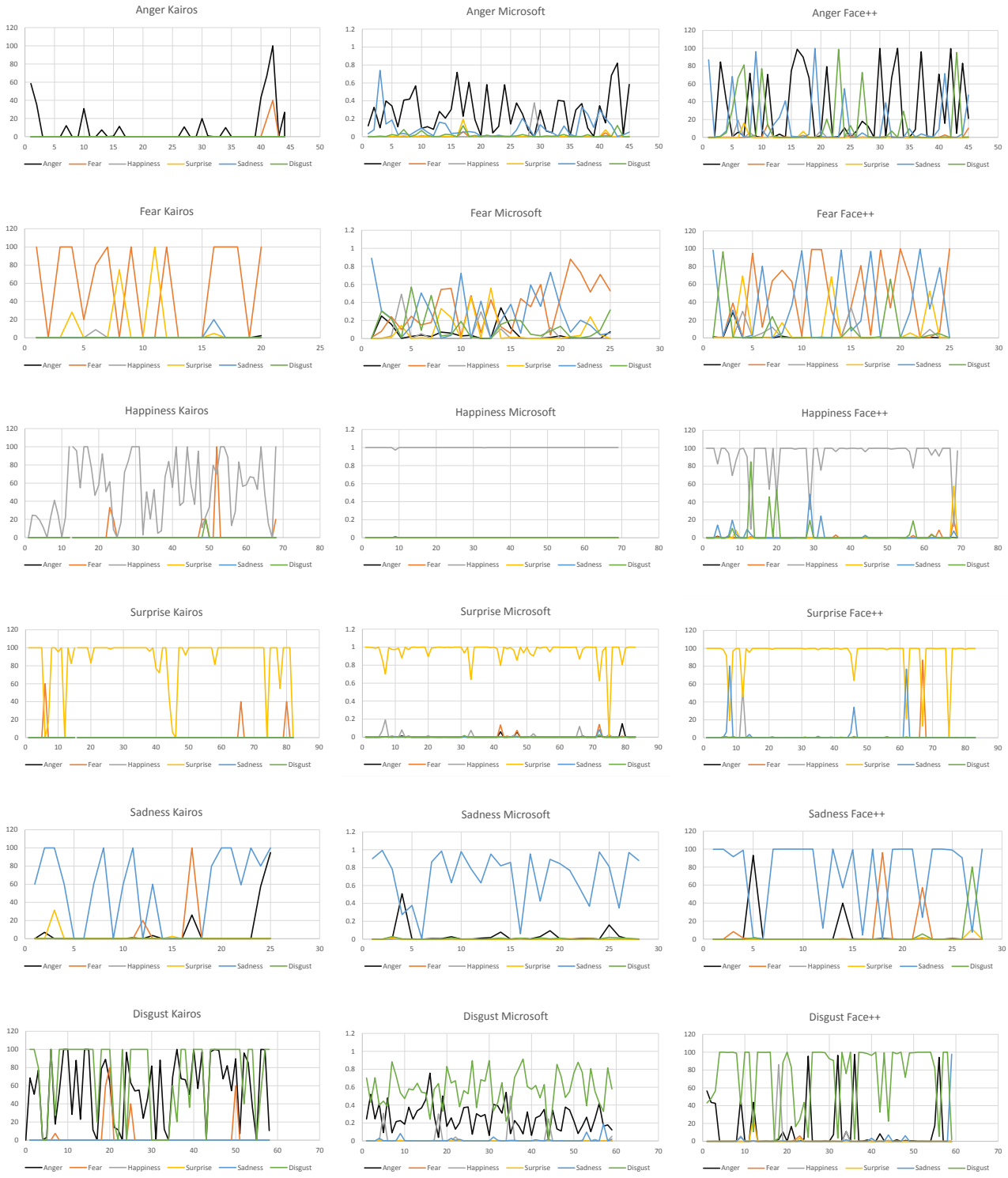

Figure 3: Output of the recognition services: recognized emotions in photos of people expressing emotions.

Table 2: Attention level: percentage correctly recognized

	Kairos	Microsoft	Face++
Covering eyes	N/A	97.37%	N/A
Covering mouth	N/A	94.74%	N/A
Covering forehead	N/A	98.68%	N/A
Closed eyes	N/A	N/A	97.37%
Attention	82.97%	N/A	N/A
Head pose attention	60.53%	N/A	72.37%
No detection: Face turned away	11.84%	7.89%	2.36%

added for which users were explicitly asked to cover part of their face. The photos were manually annotated with the features (e.g., eyes closed or not) to obtain the ground truth. The result was a dataset of 76 photos with a focus on these attention features (e.g., multiple users covering their eyes, mouth, etc.).

Table 2 shows the percentage correctly recognized photos for each attention feature. However, not all attention features are available for the three services. Features that are not available are indicated with N/A.

Face++ provides two probability values for closed eyes (for left and right eye). If both values have a probability of 40% or more, we consider this as "Closed eyes".

Kairos estimates the attention of the user and expresses this with a value between 0 (no attention at all) and 1 (full attention). Kairos attention feature is based on eye tracking and head pose. To convert this to a binary value (attention or not), we used a threshold of 0.5.

Kairos and Face++ can recognize the head pose of the user. If the head position is outside the boundaries (30 degrees for the yaw angle and 20 degrees for the pitch angle), we interpret this as "head turned away and not paying attention". The estimation of Face++ is more accurate than this of Kairos. Therefore, the head pose specified by Face++ is used in the app.

If the face is turned away too much from the camera or a large part of the face is covered, then face detection might fail. The percentage of "no detections" is also indicated in Table 2. Remember that this dataset was created with the focus on attention level. For many photos, users were explicitly asked to turn their head away. Therefore, the number of no detections is rather high.

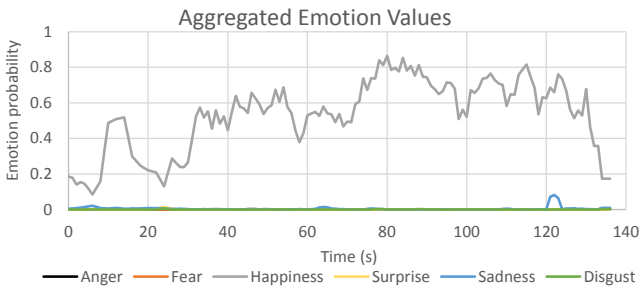

Figure 4: Emotion values aggregated over all test users.

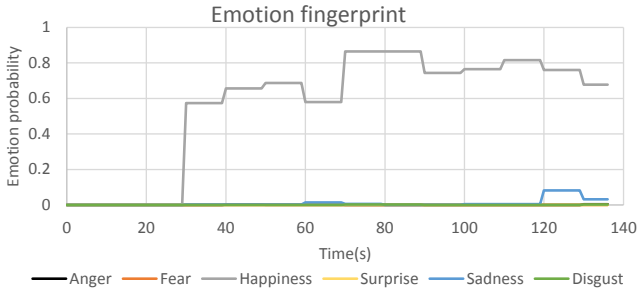

Figure 5: The emotion fingerprint based on the aggregated emotions.

4.3 Emotion recognition offline

The emotion recognition ability of the three facial recognition services was evaluated using the Cohn Kanade dataset [12, 13], which contains photos of people showing different emotions evolving from neutral to a very explicit emotion. Six photo sets with the very explicit emotions (one set for each emotion) are used as input for the facial recognition services. The output of the recognition services is a vector of 6 values, one for each emotion. For Kairos and Face++, these output values range from 0 (meaning this emotion has not been recognized at all) to 100 (meaning this emotion has been recognized with great certainty). For the Microsoft service, the output values range from 0 to 1 (with the same interpretation).

Figure 3 shows for each of the six photo sets how the emotions are recognized by the services. The emotion values are shown on the Y-axis for each photo set that was used as input (photo index on the X-axis). Each recognized emotion has a different color. For a specific photo set, the ideal emotion recognition should result in the detection of only one emotion with a value of 1 for Microsoft and 100 for Kairos and Face++, while the other emotion values are 0. For a limited number of photos, the person's face could not be detected. This resulted in no output of the service. Therefore, not all indices have an emotion value in the graphs of Kairos. In general, the results clearly show that some emotions, such as happiness and surprise, are more easy to detect with a high certainty, whereas other emotions, such as fear, are more difficult to detect and can easily be confused. Although the people of these photos are expressively showing their emotions, the automatic recognition of these emotions is not yet perfect.

Anger is accurately recognized by Kairos and Microsoft, whereas Face++ confuses anger with disgust and sadness for some photos. Fear is the most difficult to detect: Kairos detects fear in most photos; but Microsoft and Face++ sometimes incorrectly recognize sadness and disgust. Happiness is very accurately detected by all three services. With the Microsoft service, the results are almost perfect: only happiness is detected and no other emotions. Also surprise is very well recognized by all three service with high emotion values. Sadness is recognized for most photos, but in comparison to happiness and surprise, the emotion values are lower. This indicates that sadness is less clearly recognizable for emotion recognition services. Disgust is sometimes confused with anger, but Microsoft and Face++ rightly assign a much lower value to anger for most photos.

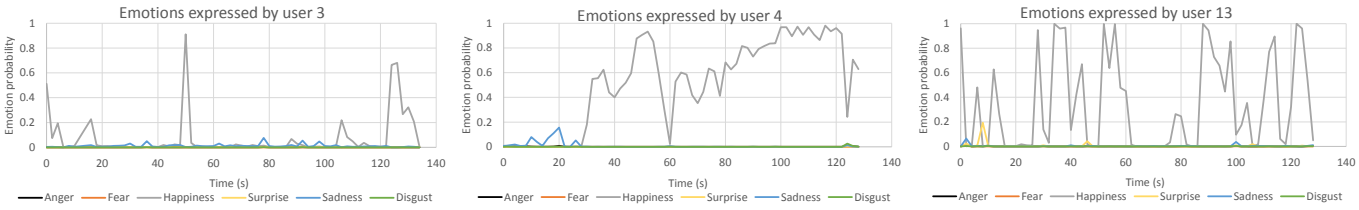

Figure 6: Emotions expressed by 3 users during video watching. Users 4 and 13 like the video, user 3 doesn't.

In conclusion, the comparison between the recognized emotions and the true emotion labels of the photos, revealed that the Microsoft service has the most accurate emotion recognition. Therefore, the Microsoft service was chosen as solution for emotion recognition in Section 4.4. The evaluation based on the Cohn Kanade dataset also indicated that - even with the most explicit emotion photos - anger, disgust, and fear are always detected with a low probability value. Happiness can be detected with high probability values. So, happiness can be considered as the emotion that is rather easy to detect with a high confidence, whereas anger, disgust, and fear are much harder to detect.

4.4 Emotion recognition online

Emotion recognition as a tool for gathering automatic feedback, was evaluated with a test panel consisting of 20 users between the ages of 5 and 72. During the test, each user watched six videos on a tablet. For each of the six basic emotions, one characteristic video was chosen (e.g., for happiness a comedy, for fear a scary horror movie, etc.). During video watching, the front-facing camera continuously took photos that were analyzed, and for which an emotion score (based on equation 1 and 2), overall attention score (equation 3), and implicit feedback score based on a complete facial analysis (equation 4) were calculated.

The emotion fingerprint of the video was obtained by aggregating the expressed emotions over all the test users. Figure 4 gives an example of this aggregation for a comedy video (a scene from the movie "Dude, Where's My Car?"). The emotion signal of the fingerprint is the average emotion value over all users at each second of the video. Because of the aggregation of emotions of multiple test persons, the emotion fingerprint was constructed after the user test. Subsequently, irrelevant emotion values are removed and only the most dominant emotions are retained (e.g., happiness, surprise, and sadness in this comedy movie). Key scenes of the video that may provoke emotions are manually selected. During periods of the video without expressive emotions, the fingerprint values are set to zero. During these periods, we assume that the emotions recognized from the users' face are due to external factors. As visible in Figure 5, the video contains no emotional scene from second 0 until 30. Next, the fluctuations of the emotion signal are reduced by using the maximum observed emotion value over a time window of 10 seconds. This takes into account that an expression of emotions typically takes multiple seconds. Figure 5 shows an example of a resulting emotion fingerprint. We consider this emotion fingerprint as the expected emotion spectrum for the specific video.

To discuss the results, we elaborate on the emotion spectrum of three users of the test. Figure 6 shows the expressed emotions of users 3, 4, and 13, while watching the comedy video. The expressed emotions of users 4 and 13 clearly show some similarities with the emotion fingerprint. Happiness is the most dominant emotion, but also some sad and surprising aspects are in the movie. The video contains the most expressive emotions (funny scene) from second 30, which is visible in the expressed emotions of user 4 and 13.

The explicit ratings for the video of users 3, 4, and 13 were respectively: 3, 6, and 6.5 stars on a scale from 1 to 10. The low explicit rating of user 3 is reflected in the emotion values of this user (implicit feedback), which are significantly lower than with the other users.

For the test with 20 users, we achieved a significant positive correlation of 0.37 between the explicit rating given by the user, and the similarity between the user's expressed emotions and the expected emotion fingerprint (equation 2). Since the rating process and emotion recognition are characterized by a lot of noise, the correlation between both will never be very high. However, the positive correlation indicates that expressed emotions clearly are a form of implicit feedback that can be used as input for a recommender system. Moreover, we expect that the correlation might improve if users watch full movies or tv shows instead of movie trailers, as in our user test. Therefore, we can consider the recognized emotions as a valid alternative feedback method in case ratings are not available, or as a feedback method 'during' content consumption instead of 'after' finishing the consumption. This answers our second research question.

Besides the emotion score, we also studied the implicit feedback score (equation 4), which is the combination of emotion and attention score. However, the variation in the attention score was limited for our user test, since all trailers are rather short (2-3 minutes). We suspect that the duration of the trailers is too short to build up intense emotional moments that make users inclined to cover their eyes or mouth. Moreover, the trailers are too short to witness a decreasing level of attention (e.g., falling asleep). Therefore, we expect that the attention score and implicit feedback score might be better suited as implicit feedback for content items with a longer duration.

5 DISCUSSION

During the user test, it became clear that people do not express their emotions much during video watching, even not if the videos contain scenes with intense emotions as selected in our test. Happiness is expressed most clearly, and is the only emotion that reached

the maximum probability value of 1, e.g., for person 13 as visible in Figure 6. For the other basic emotions, the recognition services typically register probabilities that are much lower. The second most recognizable emotion was sadness. It has a maximum value over all users of 0.68, with only 15% of the test users scoring a sadness value of 0.60 or higher (for the sad video). For fear, the maximum registered value over all test users was only 0.27 (during the fearful video). Fear is the most difficult emotion to recognize, as was also discussed in the offline test.

For this experiment, the emotion fingerprint was constructed by aggregating the emotion values of all users. A big challenge is to identify the correct expected emotions and their probability values for the fingerprint spectrum. For this, we propose the following guidelines: 1) Limit the fingerprint to a few emotions that are clearly expressed in the video. 2) Some emotions, such as fear, are more difficult to detect than others, such as happiness. The emotion probabilities from the facial recognition services are often much lower for the difficult emotions. This should be reflected in the values of the fingerprint. 3) Limit the comparison of expected and expressed emotions to the key scenes of the movie. Recognized emotions during scenes without emotions might be due to other causes than the video.

6 CONCLUSION

An Android app was developed to investigate if facial recognition services can be used as a tool for automatic authentication, user profiling, and feedback gathering during video watching. The idea is to use this feedback as input for a recommender systems. In contrast to ratings, this feedback is available during content playback. An evaluation with a test panel of 20 users showed that the authentication is almost perfect. Estimation of gender and age are in most cases accurate enough to cope with the cold-start problem by recommending movies typical for the user's age and gender. Facial analysis can be used to derive automatic feedback from the user during video watching. Closed eyes, looking away (head pose, attention level), covering eyes or mouth (occlusion), etc., are typical indications that the user does not want to see the video, and can be considered as negative implicit feedback for the recommender. By emotion recognition and a comparison with an emotion fingerprint, we calculated a user feedback value, which is positively correlated to the user's star rating. This indicates that recognized emotions can be considered as valuable implicit feedback for the recommender. Happiness can be most accurately detected. Taking photos or making videos with the front-facing camera has been expressed as a privacy-sensitive aspect by our test users and will be further tackled in future research.

REFERENCES

[1] Gediminas Adomavicius, Ramesh Sankaranarayanan, Shahana Sen, and Alexander Tuzhilin. 2005. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems (TOIS)* 23, 1 (2005), 103–145.

[2] Ioannis Arapakis, Yashar Moshfeghi, Hideo Joho, Reede Ren, David Hannah, and Joemon M Jose. 2009. Integrating facial expressions into user profiling for the improvement of a multimodal recommender system. In *2009 IEEE International Conference on Multimedia and Expo*. IEEE, 1440–1443.

[3] Microsoft Azure. 2019. Face API - Facial Recognition Software. Available at https://azure.microsoft.com/en-us/services/cognitive-services/face/.

[4] Mayank Chauhan and Mukesh Sakle. 2014. Study & analysis of different face detection techniques. *International Journal of Computer Science and Information Technologies* 5, 2 (2014), 1615–1618.

[5] Toon De Pessemier, Damien Verlee, and Luc Martens. 2016. Enhancing recommender systems for TV by face recognition. In *12th International Conference on Web Information Systems and Technologies (WEBIST 2016)*, Vol. 2. 243–250.

[6] Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology* 17, 2 (1971), 124.

[7] Michael D. Ekstrand. 2018. *The LKPY Package for Recommender Systems Experiments.* Computer Science Faculty Publications and Presentations 147. Boise State University. https://doi.org/10.18122/cs_facpubs/147/boisestate

[8] Face++. 2019. Cognitive Services - Leading Facial Recognition Technology. Available at https://www.faceplusplus.com/.

[9] IMDb. 2019. Ratings and reviews for new movies and TV shows. Available at https://www.imdb.com/.

[10] Hideo Joho, Joemon M Jose, Roberto Valenti, and Nicu Sebe. 2009. Exploiting facial expressions for affective video summarisation. In *Proceedings of the ACM international conference on image and video retrieval*. ACM, 31.

[11] Kairos. 2019. Serving Businesses with Face Recognition. Available at https://www.kairos.com/.

[12] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. 2000. Comprehensive database for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. IEEE, 46–53.

[13] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 94–101.

[14] Judith Masthoff. 2011. Group recommender systems: Combining individual models. In *Recommender systems handbook*. Springer, 677–702.

[15] Noldus. 2019. FaceReader - Facial Expression Recognition Software. Available at https://www.noldus.com/human-behavior-research/products/facereader.

[16] Commom sense media. 2019. You know your kids. We know media and tech. Together we can build a digital world where our kids can thrive. Available at https://www.commonsensemedia.org/about-us/our-mission.

[17] Marko Tkalčič, Andrej Košir, and Jurij Tasič. 2011. Affective recommender systems: the role of emotions in recommender systems. In *Joint proceedings of the RecSys 2011 Workshop on Human Decision Making in Recommender Systems (Decisions@RecSys'11) and User-Centric Evaluation of Recommender Systems and Their Interfaces-2 (UCERSTI 2) affiliated with the 5th ACM Conference on Recommender*. 9–13.

[18] Ming-Hsuan Yang, David J Kriegman, and Narendra Ahuja. 2002. Detecting faces in images: A survey. *IEEE Transactions on pattern analysis and machine intelligence* 24, 1 (2002), 34–58.