Project Proposal

2020095178 최윤선
2021029443 전예은

Topic

<Suggesting a blind date app using Object detection and Image captioning>

Why

The big goal of our project is to **"Get to know the image in a situation where we can't see the image."**

These days, many people use dating apps to find their ideal person. There are several requirements they want from future partners such as academic background, economic ability, residence, and appearance. Among these many factors, people are afraid to show their appearance by photo because someone else using the same dating app might recognize their face. At the same time, people are curious about the appearance of the person they will meet on the first day of blind date. So we are going to design some algorithms that match the future partner who has the appearance the users want and also objectively explain the appearance of the future partners.

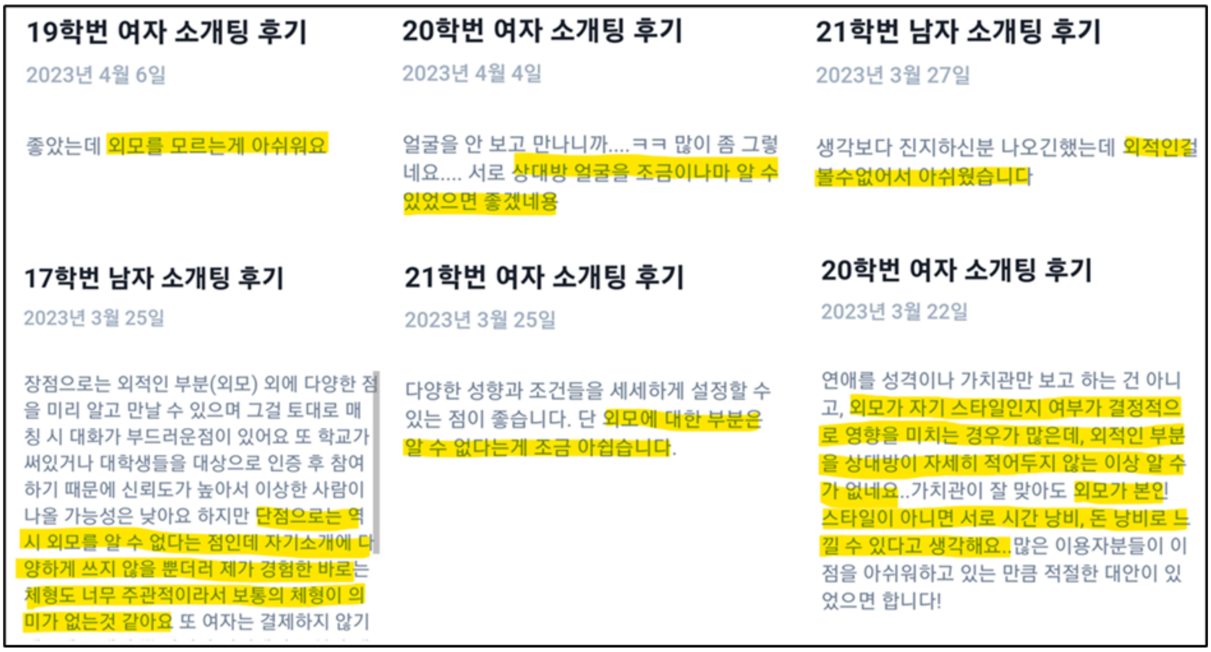
Why appearance?

It is true that we are building algorithms that focus on people's appearance, but we are not saying that appearance is the most important factor when finding a future partner. We think that appearance is just one factor that people consider when looking for a future partner among many other factors such as academic background, economic ability, residence etc. The reason that we are focusing on appearance is because other factors are easy to express by writing, but it is not easy to express appearance by writing or quantitative methods. In addition, it is difficult to describe one's face objectively on one's own (for example, skin color and eye size). Since many people have the perception that artificial intelligence is more objective than humans, if AI objectively describes people's appearance rather than people explaining their appearance by themselves, users will put more trust into the description of appearance.

Is appearance necessarily important?

Even if it is not a blind date, it is undeniable that many people are unconsciously evaluating others' appearance. Especially in the case of blind dates, appearance can be an important factor in developing feelings about future partners.

"Yeonpick" was one of the light dating apps that users cannot know much about others' appearance. In the app, people could only explain their appearance by selecting choice of some simple characteristics such as double eyelids or not. There were many reviews of the app saying that it’s too bad that they can’t know the partner’s appearance until they meet in person.



19학번 여자 소개팅 후기

2023년 4월 6일

좋았는데 외모를 모르는게 아쉬워요

20학번 여자 소개팅 후기

2023년 4월 4일

얼굴을 안 보고 만나니까....ㅋㅋ 많이 좀 그렇네요.... 서로 상대방 얼굴을 조금이나마 알 수 있었으면 좋겠네용

21학번 남자 소개팅 후기

2023년 3월 27일

생각보다 진지하신분 나오긴했는데 외적인걸 볼수없어서 아쉬웠습니다

17학번 남자 소개팅 후기

2023년 3월 25일

장점으로는 외적인 부분(외모) 외에 다양한 점을 미리 알고 만날 수 있으며 그걸 토대로 매칭 시 대화가 부드러운점이 있어요 또 학교가 써있거나 대학생들을 대상으로 인증 후 참여하기 때문에 신뢰도가 높아서 이상한 사람이 나올 가능성은 낮아요 하지만 단점으로는 역시 외모를 알 수 없다는 점인데 자기소개에 다양하게 쓰지 않을 뿐더러 제가 경험한 바로는 체형도 너무 주관적이라서 보통의 체형이 의미가 없는것 같아요 또 여자는 결제하지 않기

21학번 여자 소개팅 후기

2023년 3월 25일

다양한 성향과 조건들을 세세하게 설정할 수 있는 점이 좋습니다. 단 외모에 대한 부분은 알 수 없다는게 조금 아쉽습니다.

20학번 여자 소개팅 후기

2023년 3월 22일

연애를 성격이나 가치관만 보고 하는 건 아니고, 외모가 자기 스타일인지 여부가 결정적으로 영향을 미치는 경우가 많은데, 외적인 부분을 상대방이 자세히 적어두지 않는 이상 알 수가 없네요..가치관이 잘 맞아도 외모가 본인 스타일이 아니면 서로 시간 낭비, 돈 낭비로 느낄 수 있다고 생각해요..많은 이용자분들이 이 점을 아쉬워하고 있는 만큼 적절한 대안이 있었으면 합니다!

By these reviews, we can see that there are many people who value appearance before going on a blind date.

In a society where one is reluctant to show one’s appearance by photo, we would like to suggest a way to grasp others’ appearance without seeing the photo. We expect that users will have more trust in the description of appearance by AI’s objective explanation rather than people’s subjective explanation.

How

Our goal is to design the algorithm that matches the future partner who has the appearance the user wants and besides objectively explains the appearance of the future partner. Later on if possible, we will make a dating app including the algorithm we made.

[Algorithm] The overall process of the algorithm we are planning to make is like this.

(1) The user uploads a photo that shows the user’s appearance. We will let users know that the photo will not be released anywhere else. The other users cannot see the photo, and we have access to the photo but we will not see the users’ photo also for personal information protection.

(2) Let’s think that there are only two facial attributes(skin color, eye size) that the user can choose for a future partner. Once we get a user's photo, we classify the photo using object detection. We extract skin color from cheeks or forehead or under chin by using object detection. We set a fixed standard for brightness of skin, and then classify the photo’s skin color by the standard. (Currently we are thinking of changing the photo into grayscale and then we select some brightness value.) Also by using object detection, we can calculate the

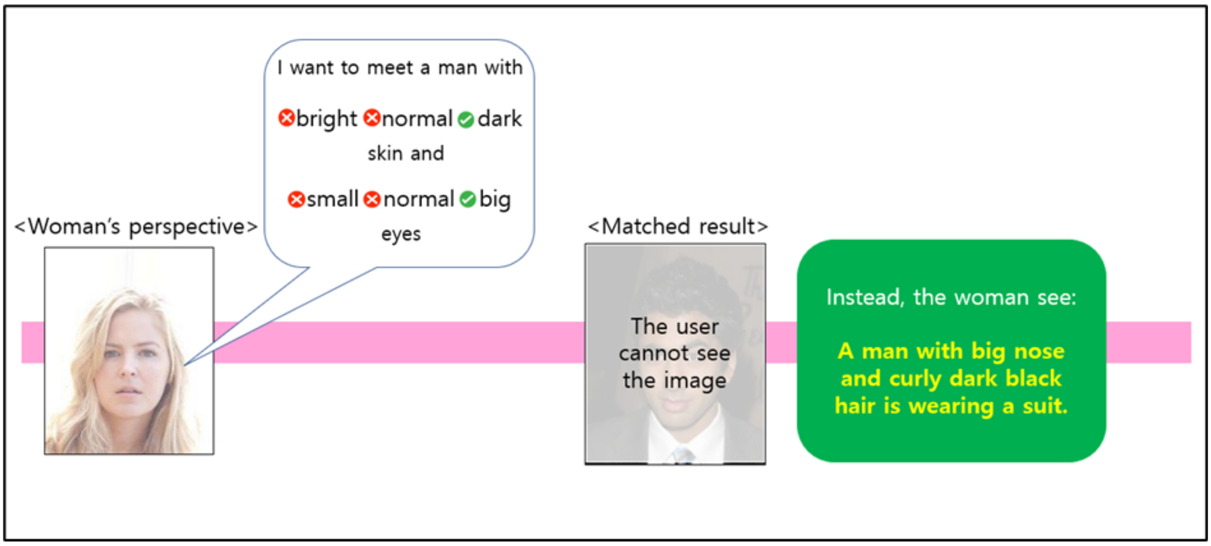
ratio of the eye area to face area. We also can set a fixed standard for the ration of the eye size, and then classify the eye size by the standard.

(3) Since we can classify these photos by object detection, we can label them with those facial attributes. By this process, we will have a huge number of users' photos with labels.

(4) Now let's go back to the user's point of view. Users will choose the facial attributes they prefer for future partners. For example, a woman with big eyes and bright skin would want to meet a man who has big eyes and dark skin. Then we will select all the men's photos that have big skin and dark eyes. Between those men, we will select men who would want to meet a woman with big eyes and bright skin. Between those selected men, we will select one random photo and match with the woman.

(5) The woman will also be curious about the additional appearance of the man. We will use image captioning to explain the man's photo. Let's say that the result of the image caption of the man's photo will be like "A man with a big nose and curly dark black hair is wearing a suit." Then we will show the image caption result to the woman, not the photo. Women can know more about the external appearance of the man by the image caption sentence and be curious about the man.

(6) We showed the whole process of matching from the woman's point of view, but the same thing happens on a man's part. If they both choose to meet each other, they meet. If one of them chooses not to meet, they don't meet and we introduce a new person by repeating the same process above.


I want to meet a man with
☒bright ☒normal ✓dark
skin and
☒small ☒normal ✓big
eyes
<Woman's perspective>
<Matched result>
The user cannot see the image
Instead, the woman see:
A man with big nose and curly dark black hair is wearing a suit.

[Dataset] We are using CelebA dataset(https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html). CelebA dataset contains 202,599 face images of the size 178×218 from 10,177 celebrities, each annotated with 40 binary labels indicating facial attributes like hair color, gender and age. We chose this dataset because people's face sizes were adjusted to the similar scale, and there were binary labels that could help us in data preprocessing.

[Data Preprocessing] Since the image has attribute annotations such as gender and heavy makeup, we will delete the images with heavy makeup for better face recognition. Then we will divide the images by gender annotation by using the library ‘to_categorical’ of tensorflow, and 5000 images will be randomly selected for each gender making 10000 images in total.

[Model] We will use object detection to match people by external ideal type, and then use image caption to specifically explain the additional external appearance. Details are in the “What” section.

What

To evaluate the features of facial images, we will perform facial landmark object detection on features such as eyes, nose, and mouth, and extract facial images that meet the users’ choices about their ideals. Among the extracted images, we will randomly select one, and perform the image captioning to generate additional descriptive sentences about the selected image beyond the chosen ideals. For these tasks, we decided to try two models for object detection and two models for image captioning.

[Object detection] Faster R-CNN & MTCNN

(1) Faster R-NN
Regarding the Faster R-CNN model, we referred to the paper "A Method of Eye and Lip Region Detection using Faster R-CNN Face Image (Jeong-Hwan Lee)" which uses Faster R-CNN to detect eye and lip regions. Since this paper shows a similar task to what our team wants to do, we aim to use Faster R-CNN to detect the locations of eyes, forehead (or chin), nose, etc. on the face, and use this information to determine the size or ratio of eyes or nose, as well as skin color.

(2) MTCNN
We plan to use MTCNN to extract facial landmarks. Especially we focus on the third step, O-Net, to perform bounding box regression for the eyes, nose, and mouth in the images. Through this, we expect to measure the size of each landmark and use them for objective facial image evaluation. To implement MTCNN, we refer to the paper 'Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks (Kaipeng Zhang)[2]'.

[Image captioning] ResNet50-LSTM & Transformer-based

(1) ResNet50-LSTM
We will implement the most commonly used CNN and RNN-based architectures for image captioning models. However, we will use ResNet50 instead of CNN and LSTM instead of RNN to train the images and texts. By using ResNet50 and LSTM, which have better performances than CNN and RNN models, we aim to get better performance of image captioning. We refer to the paper, ‘Image Caption Generator Using RESNET-LSTM (Shobiya L)[3]’.

(2) Transformer-based
We will implement the Transformer-based image captioning model, based on the paper ‘End-to-End Transformer Based Model for Image Captioning (Yiyu Wang) [4].’ Unlike the previous ResNet50-LSTM model, the Transformer-based model, which performs end-to-end

training, integrates image captioning into a single stage. Therefore, we decided to implement this model to compare with the two-stage and the one-stage image captioning model.

Who

[Together]

(1) Data preprocessing

(2) Compare and Choose the model
We selected two models for each Object detection and Image captioning tasks, so we have to compare the performance of each model to decide which one performs better. After the comparison, we will select one model for each task.

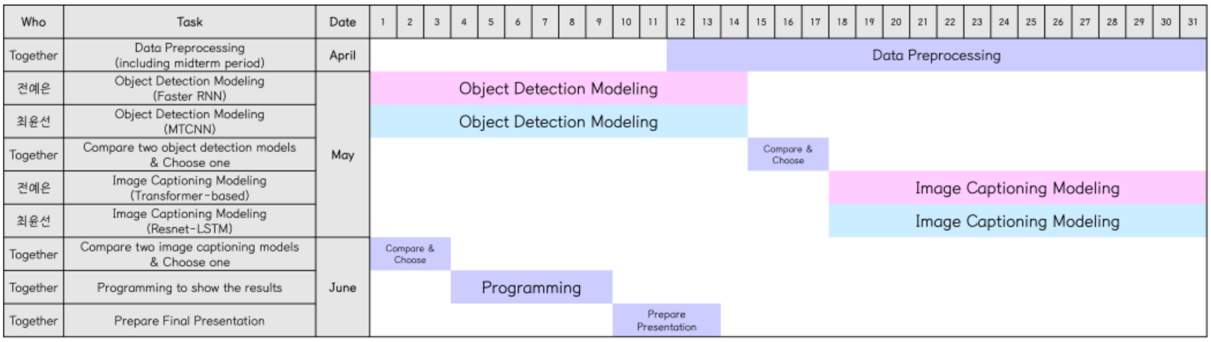
(3) Programming
We need to link the selected Object detection model code with the Image captioning model code. After that, we have to create several options related to the ideals that the user can select, derive the results that image captioning is completed for the selected image, and write a code that can be output on the screen.

[전예은] Yeeun will be implementing Faster R-CNN for Object detection, and Transformer-based Image Captioning Model.

[최윤선] Yoonseon will be implementing YOLO for Object detection, and ResNet50-LSTM based Image Captioning Model.

When

This is the outline of the project. (Gantt Chart)
https://drive.google.com/file/d/1aZWmCvzz8Xi3-ksxXR_v2wpKWh3dN_Pp/view?usp=share_link



Who	Task	Date	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Together	Data Preprocessing (including midterm period)	April												Data Preprocessing																			
전예은	Object Detection Modeling (Faster RNN)	May	Object Detection Modeling																													
최윤선	Object Detection Modeling (MTCNN)		Object Detection Modeling																													
Together	Compare two object detection models & Choose one																Compare & Choose																
전예은	Image Captioning Modeling (Transformer-based)																			Image Captioning Modeling													
최윤선	Image Captioning Modeling (Resnet-LSTM)																			Image Captioning Modeling													
Together	Compare two image captioning models & Choose one	June	Compare & Choose																													
Together	Programming to show the results					Programming																											
Together	Prepare Final Presentation											Prepare Presentation																					

Reference

[1] A Method of Eye and Lip Region Detection using Faster R-CNN Face Image
https://koreascience.kr/article/JAKO201827041051649.pdf

[2] Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks
https://arxiv.org/ftp/arxiv/papers/1604/1604.02878.pdf

[3] Image Caption Generator Using RESNET-LSTM
https://www.ijres.org/papers/Volume-9/Issue-8/Series-1/L09086471.pdf

[4] End-to-End Transformer Based Model for Image Captioning (Yiyu Wang)
https://arxiv.org/pdf/2203.15350.pdf