Person of Interest in Images

Clint Solomon Mathialagan
Virginia Tech
mclint@vt.edu

Abstract

Given a photo of a group of people, is it possible to reason about whom the photo is all about? Is there a measure that tells how important a person is in an image? If so, can the people in an image be ranked based on their relative importance? This project aims to find answers to the above questions.

1. Introduction

The objective of this work is to determine a measure for importance of people in images. What is importance? While there is no absolute definition for importance in this context, the idea of importance can be elucidated with the following properties. Importance of a person in an image relates to how likely it is to find the person mentioned in a sentence describing the image. Importance can also give an idea of how distinguished a person is in an image. It can also be considered as an indicator of the main character(s) in an image.

1.1. Motivation

To consider the benefits of studying importance, applications of such an importance measure are considered. In social networking sites, when a query comes in for photos of a particular person, there can be a random ranking of the person's photos or date based ranking or some other ranking scheme that is not dependent on the characteristics of the photo that is relevant to the person. If there exists a measure of importance of the person in every image, one better way to service this query would be to return a sorting of the photos based on the person's importance in these images.

In context-aware image resizing techniques such as seam carving, one limitation is that image regions with people tend to get distorted badly. This can be fixed by applying high potentials on such regions. However if there are many people present in the image, such a technique will affect resizing capabilities. With the proposed measure of importance, the potentials for these regions can be made proportional to the importance of the person in that region and thereby the main characters can be retained without affecting resizing capability.

In natural language processing, given an image with a number of people in it, most existing techniques would generate a sentence for each person. Using an importance measure can help reduce the number of sentences by generating sentences only for the person(s) of interest in these images. Thus it would be a more natural way of describing such a scene.

1.2. Challenges

There is an inherent ambiguity when reasoning about importance. Given two photos as in Figure 1, it may be unambiguous to identify the main character in the photo to the left whereas the photo on the right does not have a clear "important" character nor a clear distinguished person. What should an ideal importance measure be in this case? The measure should neither be low nor be high for any person in this image. However if three of the persons in this photo moved to the background and were involved in some irrelevant task, the ideal importance measure should be high for the fourth person whose position is retained. This points to the fact that the measure of importance depends not only on the characteristics of that person in the image but also on the other people in the image. Thus the importance measure should capture both global and local information. Another subtle point to note is that in many cases the order of importance may not be a strict order or may only be significant in the top ranking elements. Hence the learning method should not be penalized for minor perturbations in the bottom of the order.

Figure 1. Amibuity in Importance

1.3. Previous Work

This work is novel and there is no previous work in this exact nature. Some of the closest works involve Andrew Gallagher's work [1] on images of people.

2. Dataset

Since the task here is novel, there is no standard dataset suitable for this learning objective. Most of the image datasets either have images with few people in them or they do not have images that can serve as good samples for importance measures. The ideal dataset should have images with characteristics as shown in Figure 1 i.e they should have images with clearly defined important people as well as ones with no such clear demarcation. The collected dataset has 500 images from sources explained as below.

2.1. The Images Of Groups Dataset

This is a dataset [3] used by Andrew Gallagher for his work. The dataset contains images of people posing to the camers in various occasions. The people are all facing the cameras with eyes mostly parallel to the horizontal of the image. The dataset comes with annotated faces, eyes, mouth and noses and calculated roll, pitch, yaw etc.This dataset has 5080 images of groups of people in weddings, family potraits and general shots. These images were mined from Flickr using specific search keywords. However since a majority of these photos have people posing to camera, there is not much of importance-related information in these images meaning these images fall into the ambigious category explained before. Out of the 500 images used for this report, 250 were collected from this dataset.

Figure 2. Example from Images of Groups Dataset

2.2. Mining Flickr

For images that can have more definite importance measures, the Flickr download script by Tamara Berg and James Hays was adapted. This script was originally written to download geo-tagged images and it searches for keywords in any text related to the image. This was modified to search only in tags and also to return only those images that have the relevant license that would allow their use in academic research. Another 250 images were mined this way to create the entire dataset.

Figure 3. Examples of images added from Flickr

3. Annotation

The dataset then required both feature as well as importance to be annotated. The cost incurred for annotation was borne by Dr.Dhruv Batra of the Machine Learning and Perception Lab at Virginia Tech.

3.1. Feature Annotation

Before annotating the images in Amazon Mechanical Turk, the images had to be annotated with faces so that mturk workers can be asked to click on the faces to select a person. Also the features used in the learning required eye locations to identify the two dimensional orientation of the face. These annotations are present in the Images of Groups dataset. The 250 images downloaded from Flickr were then annotated using an API provided by Sky Biometry. The API was chosen because it gave a very low false positive rate (one non-maximal suppression issue in 250 runs). Ideally a state-of-the-art algorithm should be used. Since this is a proof-of-concept endeavor, the API was preferred. The API also returns several other attributes like gender, age, smiling, glasses etc. These are planned to be used in future. The faces missed by this API were then manually added by the author. Also it should be noted that at this stage, the annotation was done so that any person whose head was visible was annotated to the best of ability. This was done because the focus is on the person and not just on faces. Hence these regions will have missing eye, mouth or nose annotations. The missing eyes were taken care of in the 2D orientation feature extraction.

3.2. Importance Annotation

The images were then annotated on Amazon Mechanical Turk by asking the mturk workers to select people of interest in the images. In the first batch, 350 images were annotated by running 70 HITs with 5 images each. These were priced at 3 cents a HIT. The number of assignments were chosen to be 20 so that the annotation can overcome the noise owing to the personal preferences of the mturk workers. The task ran very slow and had to be pulled out.

The second annotation task had the remaining images with 10 images ber HIT and 5 cents per HIT. This was also slow and was still running at the time this report was written. Thus the current results were obtained with 350 annotated images. Figure 4 shows a screenshot of the annotation interface.

Figure 4. Mturk Annotation Interface

4. Feature Extraction

The features used are listed in Table 1.When calculating orientation, if eyes are missing, the values of the orientaton are shifted appropriately (to discrete values), assuming that a person is showing the right part of his face if his left eye is missing and so on. Thus this measure is not strictly two dimensional and is not a complete three dimensional measure as well. One point to be noted here is that the scale feature depends to an extent on how the annotation was done. Manual annoations tend to have larger areas and these may in some way also encode that the person was tougher to identify automatically.

Feature	Description
Centerness	Normalized distance of the person from the center of the image
Orientation	Two dimensional orientation of the person computed from eye co-ordinates
Scale	Scale of the bounding box of the person's head/face normalized by image dimensions
Sharpness	Normalized count of the number of edge pixels in the bounding box

Table 1. Features used

5. Learning Importance

SVM classifiers and rankers were used in learning importance from these images. The idea was to see if ranking can be done reasonably with just the classifier's score. Another motivation was to see if the global features of a person is good enough to provide a measure of importance in any image.

5.1. Classifier

The classifier was trained on data for a 1000 people and tested on 1143 people. Care was taken to ensure a good amount of positive and negative samples in both test and training data. A validation set containing data for almost a 100 people were used to choose some of the parameters but this was not rigorously tested as with the inclusion of the pending data from mturk these would have to be recalculated anyways. The plan here was to get a ball park for accuracies and for the kind of methods that would finally make sense for this task.

The libsvm [2] package was used to create variants of SVM classifiers. Third order polynomial, linear, RBF and Sigmoid kernels were compared using a normalized accuracy measure which is the average of true positive rates and false positive rates. The baselines included an always true predictor, an always false predictor and a predictor that takes as input the number of positives needed and selects that many numbers based on the distance of the person from the center of the respective image. For the result shown in Figure 5 , the actual number of positives in the image was used.

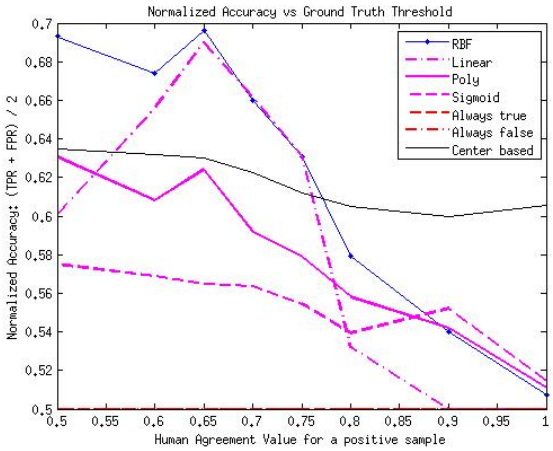

Figure 5. Comparision of performance of different SVM kernels

On the X-axis is shown the value at which a person is considered to be important. A value of 0.75 would indicate that 75 percent of the annotators thought that the person was important. Thus the curves are obtained by varying the threshold at which ground truth is generated. It is found that the RBF method works the best when the ground truth is thresholded at 75 percent or lesser. As this threshold is increased, the classifier performance drops and the center based baseline performs better though the best it does is in the 0.62 accuracy region. The best the classifier performs

is around 0.69. This shows that there is plenty of room for improvement and that it is possible to beat the tough center baseline.

In figure 6, the performance (prediction accuracy - unnormalized) of the various features is shown with increasing slack penalty. The RBF kernel with 0.75 ground truth threshold was chosen for these experiments. The centerness feature performs the best individually and also beats all the other features combined together. The best performance is obtained when all the features are combined together. This shows that these simple features are good starting points. It is seen that once the slack penalty is at a sufficient value, the performance is almost constant to its increase beyond that.

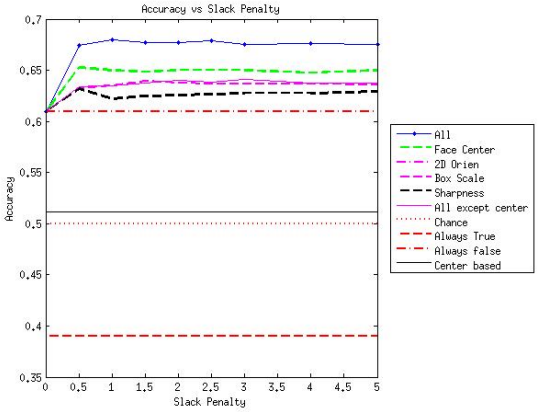

Figure 6. Comparision of performance of different features

5.2. Ranker

The SVM ranker [4] provided by Joachims is used to see how a ranking method would perform at this task. For this analysis, the ground truth is the ranking of people in an image based on the annotated importance values. These were fed as constraints to the rank SVM and the performance was evaluated using Kendall's Tau and ranking Prediction Error measures. The classifier scores were also used to generate rankings and this was compared against those produced by the ranker. The baselines include a random rank predictor that creates a random ranking given an image and a center based predictor that ranks the person nearest to the center first and so on. The table 2 summarizes the result of this experiment. It is found that the ranker outperforms all the rest.

6. Conclusion

From the experiments conducted, it is observed that there is lot of scope for improvement in this work. The fact that the classifiers are able to beat the simple center based predictor is promising as this seemed to be a very tough baseline. The fact that the ranker beats the classifier gives the

Method	Kendall's Tau	Prediction Error
Rank SVM (c=5)	0.3460	0.3270
SVM Classifier (c=5)	0.2479	0.3761
Center Based Ranking	0.2224	0.3888
Random Ranking	0.0133	0.4933

Table 2. Comparision Results

implication that importance should be learnt from information of other people in the image in addition to the features of the person. Maybe a probabilistic graphical model can help. The simple features seem to be good choices but the feature set can be enriched with semantic features as well. The dataset should also be increased and care should be taken to avoid biases in the data collection as well as in the annotation process. Future work will be along these lines and hopefully accuracy can be improved enough to make this idea practical.

References

[1] T. C. Amir Sadovnik, Andrew Gallagher. Not everybodys special: Using neighbors in referring expressions with uncertain attributes. 2

[2] C.-J. L. Chih-Chung Chang. Libsvm. http://www.csie.ntu.edu.tw/ cjlin/libsvm/. 3

[3] A. Gallagher. Images of groups. http://chenlab.ece.cornell.edu/people/Andy/groupsOfPeople.html. 2

[4] T. Joachims. Svmlite rank. http://svmlight.joachims.org/. 4