VIP: Finding Important People in Images

Clint Solomon Mathialagan
Virginia Tech

Andrew C. Gallagher
Google Inc.

Dhruv Batra
Virginia Tech

Project: https://computing.ece.vt.edu/~mclint/vip/ Demo: http://cloudcv.org/vip/

Abstract

People preserve memories of events such as birthdays, weddings, or vacations by capturing photos, often depicting groups of people. Invariably, some individuals in the image are more important than others given the context of the event. This paper analyzes the concept of the importance of individuals in group photographs. We address two specific questions – Given an image, who are the most important individuals in it? Given multiple images of a person, which image depicts the person in the most important role? We introduce a measure of importance *of people in images and investigate the correlation between importance and visual saliency. We find that not only can we automatically predict the importance of people from purely visual cues, incorporating this predicted importance results in significant improvement in applications such as im2text (generating sentences that describe images of groups of people).*

1. Introduction

When multiple people are present in a photograph, there is usually a story behind the situation that brought them together: a concert, a wedding, a presidential swearing-in ceremony (Fig. 1), or just a gathering of a group of friends. In this story, not everyone plays an equal part. Some person(s) are the main character(s) and play a more central role.

Consider the picture in Fig. 2a. Here, the important characters are the couple who appear to be the British Queen and the Lord Mayor. Notice that their identities and social status play a role in establishing their positions as the key characters in that image. However, it is clear that even someone unfamiliar with the oddities and eccentricities of the British Monarchy, who simply views this as a picture of an elderly woman and a gentleman in costume receiving attention from a crowd, would consider those two to be central characters in that scene.

Fig. 2b shows an example with people who do not appear to be celebrities. We can see that two people in foreground are clearly the focus of attention, and two others in the background are not. Fig. 2c shows a common group photograph, where everyone is nearly equally important. It is clear that even without recognizing the identities of people, we as humans have a remarkable ability to understand social roles and identify important players.

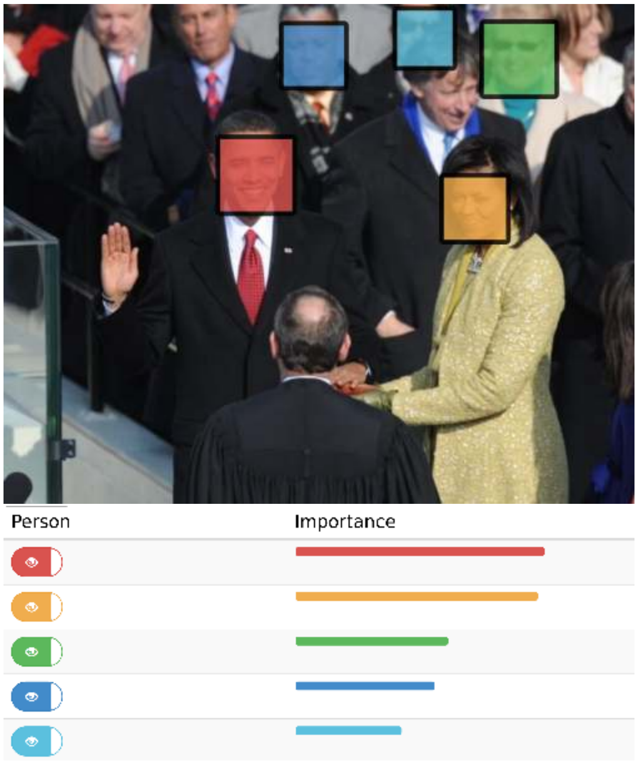
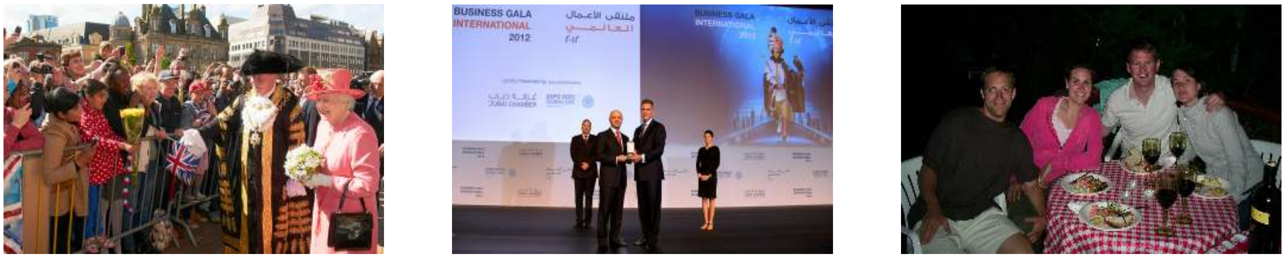


Figure 1: Goal: Predict the importance of individuals in group photographs (without assuming knowledge about their identities).

Goal and Overview. The goal of our work is to *automatically predict the importance of individuals in group photographs*. In order to keep our approach general and applicable to any input image, we focus purely on visual cues available in the image, and do not assume identification of the individuals. Thus, we do not use social prominence cues. For example, in Fig. 2a, we want an algorithm that identifies the elderly woman and the gentleman as the top-2 most important people that image without utilizing the knowledge that the elderly woman is the British Queen.



(a) Socially prominent people.

(b) Non-celebrities.

(c) Equally important people.

Figure 2: Who are most important individuals in these pictures? (a) the couple (the British Queen and the Lord Mayor); (b) the person giving the award and the person receiving it play the main role; (c) everyone seems to be nearly equally important. Humans have a remarkable ability to understand social roles and identify important players, even without knowing identities of the people in the images.

What is Importance? In defining importance, we can consider the perspective of three parties (which may disagree):

- **the photographer**, who presumably intended to capture some subset of people, and perhaps had no choice but to capture others;
- **the subjects**, who presumably arranged themselves following social inter-personal rules; and
- **neutral third-party human observers**, who may be unfamiliar with the subjects of the photo and the photographer's intent, but may still agree on the (relative) importance of people.

Navigating this landscape of perspectives involves many complex social relationships: the social status of each person in the image (an award winner, a speaker, the President), and the social biases of the photographer and the viewer (*e.g*., gender or racial biases); many of these can not be easily mined from the photo itself. At its core, the question itself is subjective: if the British Queen "photo-bombs" while you are taking a picture of your friend, is she still the most important person in that photo?

In this work, to establish a quantitative protocol, we rely on the wisdom of the crowd to estimate the "ground-truth" importance of a person in an image. We found the design of the annotation task and the interface to be particularly important, and discuss these details in the paper.

Applications. A number of applications can benefit from knowing the importance of people. Algorithms for im2text (generating sentences that describe an image) can be made more human-like if they describe only the important people in the image and ignore unimportant ones. Photo cropping algorithms can do "smart-cropping" of images of people by keeping only the important people. Social networking sites and image search applications can benefit from improving the ranking of photos where the queried person is important.

Contributions. This paper makes the following contributions. First, we learn a model for predicting importance of individuals in photos based on a variety of features that capture the pose and arrangement of the people. Second, we collect two importance datasets that serve to evaluate our approach, and will be broadly useful to others in the community studying related problems. Finally, we show that we can automatically predict the importance of people with high accuracy, and incorporating this predicted importance in applications such as im2text leads to significant improvement. Despite the naturalness of the task, to the best of our knowledge, this is the first paper to directly infer the importance of individuals in the context of a single group image.

2. Related Work

General Object Importance. Our work is related to a number of previous works [4, 13, 23] that study the importance of generic object categories. Berg *et al*. [4] define importance of an object as the likelihood that it will be mentioned in a sentence written by a person describing the image. The key distinction between their work and ours is that they study the problems at a category level ("are people more important than dogs?"), while we study it at an instance level, restricted to instances of people ("is person A more important than person B in this image?"). One result from [4] is that 'person' generally tends to be the most important category. Differentiating between the importance of different individuals in an image produces a more fine-grained understanding of the image. Le *et al*. [16] consider people who have appeared repeatedly in a certain time period from large news video databases to be important. Lee *et al*. [17] study importance of objects (including people) in egocentric videos, where important things are those with which the camera wearer has significant interaction. In our work, we focus on a single image, and do not assume access to user-attention cues.

Visual Saliency. A number of works [6, 11, 19] have studied visual saliency – identifying which parts of an image draw viewer attention. Humans tend to be a naturally salient content in images. Jiang *et al*. [14] study visual saliency in group photographs and crowded scenes. Their objective is to build a visual saliency model that takes into account the presence of faces in the image. Although they study the same content as our work (group photographs), the goals of the two are different – saliency vs importance. At a high level, saliency is about what draws the viewer's attention; importance is a higher-level concept about social roles. We conduct extensive human studies and discuss this comparison in the paper. Saliency is correlated to, but not identical to importance. People in photos may be salient but not important, important but not salient, both, and neither.

Understanding Group Photos. Our work is related to a line of work in Computer Vision studying photographs of groups of people [7–9, 20, 21], addressing issues such structural formation and attributes of groups. Li *et al*. [18] predict the aesthetics of a group photo. If the measure is below a threshold, photo cropping is suggested by eliminating unimportant faces and regions that do not seem to fit in with the general structure of the group. While their goal is closely related to ours, they study *aesthetics*, not importance. To the best of our knowledge, this is the first work to predict importance of individuals in a group photo.

3. Approach

Recall that our goal is to model and predict the importance of people in images. We model importance in two ways:

- **Image-Level Importance:** "Given an image, who is the most important individual?" This reasoning is local to the image in question. The objective is to predict an importance score for each person in the image.
- **Corpus-Level Importance:** "Given multiple images, in which image is a specific person most important?" This reasoning is across a corpus of photos (each containing a person of interest), and the objective is to assign an importance score to each image.

3.1. Dataset Collection

For each setting, we curated and annotated a dataset.

Image-Level Dataset. In this setting, we need a dataset of images containing at least three people with varying levels of importance. While the 'Images of Groups' dataset [7] initially seems like a good candidate, it is not suitable for studying importance because there is little change in *relative* importance – most images are posed group photos where everyone is nearly equally important (*e.g*. Fig. 2c).

We collected a dataset of 200 images by mining Flickr for images (with appropriate licenses) using search queries such as "people+events", "gathering", *etc*. Each image has three or more people in varying levels of importance. In order to automatically predict the importance of individuals in the image, they need to be detected first. For the scope of this work, we assume face detection to be a solved problem. Specifically, the images in our dataset were first run through a face detection API [22], which has a fairly low false positive rate. Missing faces and heads were then annotated manually. There are in total 1315 annotated people in the dataset, with ∼6.5 persons per image on average. Example images are shown throughout the paper and the dataset is publicly available from the project webpage [2].

Corpus-Level Dataset. In this setting, we need a dataset that has multiple pictures of the same person; and multiple sets of such photos. The ideal source for such a dataset are social networking sites. However, privacy concerns hinder the annotation of these images via crowdsourcing. TV series, on the other hand, have multiple frames with the same people and are good sources to obtain such a dataset. Since temporally-close frames tend to be visually similar, these videos should be properly sampled to get diverse images.

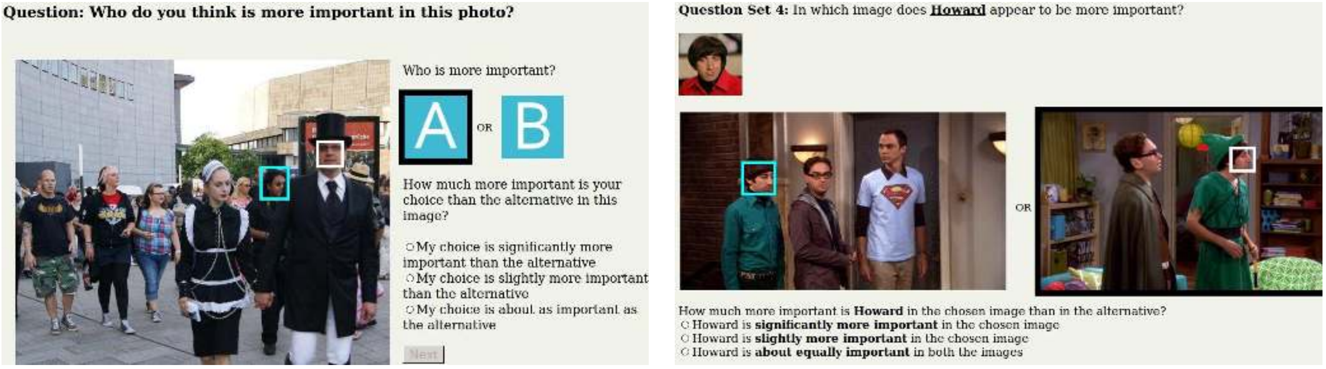
The personID dataset by Tapaswi *et al*. [24] contains face track annotations (with character identification) for the first six episodes of the 'Big Bang Theory' TV series. The track annotation of a person gives the coordinates of face bounding boxes for the person in every frame. By selecting only one frame from each track of a character, one can get diverse frames for that character from the same episode. From each track, we selected the frame that has the most people. Some selected frames have only one person in them, but that is acceptable since the task is to pick the most important frame for a person. In this manner, a distinct set of frames was obtained for each of the five main characters in each episode.

3.2. Importance Annotation

We collected ground-truth importance in both datasets via Amazon Mechanical Turk (AMT). We conducted pilot experiments to identify the best way to annotate these datasets, and pose the question of importance. We found that when subjects were posed an absolute question "Please mark the important people in this image," they found the task difficult. Turkers commented that they had to redefine their notion of importance for each new image, making consistency difficult. Indeed, we observed low inter-human agreement, with some workers selecting everyone in the image as important, and others selecting no more than one person.

To overcome these inconsistencies, we redesigned the tasks to be relative (details next). This made each task simpler, and the annotations more consistent.

Image-Level Importance Annotation. From each image in the image-level dataset, random pairs of faces were selected to produce a set of 1078 pairs. These pairs cover 91.82% of the total faces in these images. For each selected pair, ten AMT workers were asked to pick the more important of the two. The interface is shown in Fig. 3a, and an HTML version is available from the project webpage [2]. In addition to clicking on the more important face, the workers were also asked to report magnitude of the difference in importance between the two people: *significantly different*, *slightly different* and *almost same*. This forms a three-tier scoring system as depicted in Table 1.



Question: Who do you think is more important in this photo?

Who is more important?

How much more important is your choice than the alternative in this image?

○ My choice is significantly more important than the alternative
○ My choice is slightly more important than the alternative
○ My choice is about as important as the alternative

(a) Image-Level annotation interface.

Question Set 4: In which image does Howard appear to be more important?

How much more important is **Howard** in the chosen image than in the alternative?
○ Howard is **significantly more important** in the chosen image
○ Howard is **slightly more important** in the chosen image
○ Howard is **about equally important** in both the images

(b) Corpus-Level annotation interface.

Figure 3: Annotation Interfaces used with MTurk: (a) Image-Level: Hovering over a button (A or B) highlights the person associated with it (b) Corpus-Level: Hovering over a frame shows the where the person is located in the frame.

Turker selection: A is	A's score	B's score
significantly more important than B	1.00	0.00
slightly more important than B	0.75	0.25
about as important as B	0.50	0.50

Table 1: Converting pairwise annotations to importance scores.

Pair category	Image-Level	Corpus-Level
significantly-more	32.65%	18.30%
slightly-more	20.41%	39.70%
almost-same	46.94%	42.00%

Table 2: Distribution of Pairs in the Datasets.

For each annotated pair of faces (p_i, p_j) the relative importance scores s_i and s_j range from 0 to +1, and indicates the relative difference in importance between p_i and p_j. Note that s_i and s_j are not absolute, as they are not calibrated for comparison to another person, say p_k from another pair.

Corpus-Level Importance Annotation. From the corpus-level dataset, approximately 1000 pairs of frames were selected. Each pair contains frames depicting the same person but from different episodes. This ensures that the pairs do not contain similar looking images. AMT workers were shown a pair of frames for a character and asked to pick the frame where the character appears to be more important. The interface used is as shown in Fig. 3b, and an HTML version is available from the project webpage [2].

Similar to the previous setting, workers were asked to pick a frame and indicate the magnitude of difference in importance of the character. These qualitative magnitude choices were converted into scores as in shown Table 1.

Table 2 shows a breakdown of both datasets along the magnitude of differences in importance. We note some interesting similarities and differences. Both datasets have nearly the same percentage of pairs that are 'almost-same'. The instance-level dataset has many more pairs in the 'significantly-more' category than the corpus-level dataset. This is because in a TV series dataset, the characters in a scene are usually playing some sort of a role in the scene, unlike typical consumer photographs that tend to contain many people in the background. Overall, both datasets contain a good mix of the three categories.

3.3. Importance Model

We now formulate a relative importance prediction model that is applicable to both tasks: image-level and corpus-level. As we can see from the dataset characteristics in Table 2, our model should not only be able to say which person is more important, but also predict the relative strengths between pairs of people/images. Thus, we formulate this as a regression problem. Specifically, given a pair of people (p_i, p_j) (coming from the same or different images) with scores s_i, s_j, the objective is to build a model M that regresses to the difference in ground truth importance score:

$$M(p_i, p_j) \approx S_i - S_j \tag{1}$$

We use a linear model: $M(p_i, p_j) = \mathbf{w}^\intercal \boldsymbol{\phi}(p_i, p_j)$, where $\boldsymbol{\phi}(p_i, p_j)$ are the features extracted for this pair, and $\mathbf{w}$ are the regressor weights. We use ν-Support Vector Regression to learn these weights. Our pairwise feature $\boldsymbol{\phi}(p_i, p_j)$ are composed from features extracted for individual people $\boldsymbol{\phi}(p_i)$ and $\boldsymbol{\phi}(p_j)$. In our preliminary experiments, we compared two ways of composing these individual face features – using difference of features $\boldsymbol{\phi}(p_i, p_j) = \boldsymbol{\phi}(p_i) - \boldsymbol{\phi}(p_j)$; and concatenating the two individual features $\boldsymbol{\phi}(p_i, p_j) = [\boldsymbol{\phi}(p_i); \boldsymbol{\phi}(p_j)]$. We found difference of features to work better, and all results in this paper are reported with that.

3.4. Person Features

We now describe the features we used to assess importance of a person. Recall that we assume that faces in the images have been detected (by running a standard face detector).

Distance Features. We use a number of different ways to capture distances between faces in the image.

Photographers often frame their subjects. In fact, a number of previous works [5, 25, 26] have reported a "center bias" – the objects or people closest to the center tend to be the most important. Thus, we first scale the image to a size of (1, 1), and compute two distance features:
Distance from center: The distance from the center of the face bounding box to the center of the image (0.5, 0.5).

Weighted distance from center: The previous feature divided by the largest dimension of the face box, so that larger faces are not considered to be farther from the center.

We compute two more features to capture how far a person is from the center of a group:
Normalized distance from centroid: First, we find the centroid of all the center points of the face boxes. Then, we compute the distance of a face to this centroid.

Normalized distance from weighted centroid: Here, the centroid is calculated as the weighted average of center points of faces, the weight of a face being the ratio of the area of the head to the total area of faces in the image.

Scale. Large faces in the image often correspond to people who are closer to the camera, and perhaps more important. This feature is a ratio of the area of the face bounding box to the the area of the image.

Sharpness. Photographers often use a narrow depth-of-field to keep the indented subjects in focus, while blurring the background. In order to capture this phenomenon, we compute a sharpness feature in every face. We apply a Sobel filter on the image and compute the the sum of the gradient energy in a face bounding box, normalized by the sum of the gradient energy in all the bounding boxes in the image.

Face Pose Features. The facial pose of a person can be a good indicator of their importance, because important people often tend to be looking directly at the camera.

DPM face pose features: We resize the face bounding box patch from the image to 128×128 pixels, and run the face pose and landmark estimation algorithm of Zhu *et al*. [28]. Note that [28] is mixture model where each component corresponds to a an the angle of orientation of the face, in the range of -90° to +90° in steps of 15°. Our pose feature is this component id, which can range from 1 to 13. We also use a 13-dimensional indicator feature that has a 1 in the component with maximum score and zeros elsewhere.

Aspect ratio: We also use the aspect ratio of the face bounding box is as a feature. While the aspect ratio of a face is typically 1:1, this ratio can differentiate between some head poses such as frontal and lateral poses.

DPM face pose difference: It is often useful to know where a crowd is looking and where a particular person is looking. To capture this pose difference between a person and others, we compute the pose of the person subtracted by the average pose of every other person in the image, as a feature.

Face Occlusion. Unimportant people are often occluded by others in the photo. Thus, we extract features to indicate whether a face might be occluded.

DPM face scores: We use the difficulty in being detected as a proxy for occlusion. Specifically, we use scores for each the 13 components in the face detection model of [28] as a feature. We also use the score of the dominant component.

Face detection success: This is a binary feature indicating whether the face detection API [22] we used was successful in detection the face, or whether it required human annotation. The API achieved a nearly zero false positive rate on our dataset. Thus, this feature served a proxy for occlusion since that is where the API usually failed. Note that this feature requires human inspection and would not be available to a fully-automatic approach. An online demo of our system available at [1, 3] does not use this feature.

In total, we extracted 45 dimensional features for every face.

4. Results

For both datasets, we perform cross-validation on the annotated pairs. Specifically, we split the annotated pairs into 10 folds. We train the SVRs on 8 folds, pick hyper-parameters (C in the SVR) on 1 validation fold, and make predictions on 1 test fold. This process is repeated for each test fold, and we report the average across all 10 test folds.

Baselines. We compare our proposed approach to three natural baselines: center, scale, and sharpness baselines, where the person closer to the center, larger, or more in focus (respectively) is considered more important. The center baseline uses the weighted distance from center which not only gives priority to distance from the center but also to the size of the face. In order to measure how well a saliency detector performs on the importance prediction task, we used the method of Harel *et al*. [10, 12] to produce saliency maps and computed the fraction of saliency intensities inside each face as a measure of its importance.

We measure inter-human agreement in a leave-one-human-out manner. In each iteration, responses of nine workers are averaged to get the ground-truth, and the response of the tenth worker is evaluated as the human response. This is then repeated for all ten human responses and the average is reported as inter-human agreement. In order to keep all automatic methods comparable to these inter-human results, we train all methods ten times, once for each leave-one-human-out ground-truth, and report the average results.

Metrics. We use mean squared error to measure the performance of our relative importance regressors. In addition, we convert the regressor output into binary classification by thresholding against zero. For each pair of faces (p_i, p_j), we use a weighted classification accuracy measure, where the weight is the ground-truth importance score of the more important of the two, *i.e.* $\max\{s_i, s_j\}$. Notice that this metric cares about the correct classification of 'significantly-more' pairs more than the other pairs, which is natural.

Image-Level Importance Results. Table 3 shows the results for different methods. We can see that the best baseline achieves 89.55% weighted accuracy, whereas our approach achieves 92.72%. Overall, we achieve an improvement of 3.17% (3.54% relative improvement). The mean squared error for our SVR is 0.1489.

Method	Weighted accuracy
Inter-human agreement	$96.68 \pm 0.40\%$
Our approach	$\mathbf{92.72} \pm 0.93\%$
Saliency detector	$83.52 \pm 1.29\%$
Center baseline	$89.55 \pm 1.12\%$
Scale baseline	$88.46 \pm 1.13\%$
Sharpness baseline	$87.45 \pm 1.20\%$

Table 3: Image-Level: Performance compared to baselines.

Table 4 show a break-down of the accuracies into the three categories of annotations. We can see that our approach outperforms the strongest baseline (Center) in every category, and the largest difference happens in the 'significantly-more' category, which is quite useful.

Pair category	Ours	C-Baseline	Improvement
significantly-more	94.66%	86.65%	8.01%
slightly-more	78.80%	76.36%	2.44%
almost-same	55.98%	52.96%	3.02%

Table 4: Image-Level: Category-wise distribution of our predictions compared to Center baseline.

Fig. 4 shows some qualitative results. We can see that individual features such center, sharpness, scale, and face occlusion help in different cases. In 3(c), the woman in blue is judged to be the most important, presumably because she is a bride. Unfortunately, our approach does not contain any features that can pick up on such social roles.

Corpus-Level Importance Results. Table 5 shows the results for the corpus-level experiments. Interestingly, the strongest baseline in this setting is sharpness, rather than the center. This makes sense since the dataset is derived from professional videos; the important person is more likely to in focus compared to others. Our approach outperforms all baselines, with an improvement of 4.18% (4.72% relative improvement). The mean squared error is 0.1078.

Table 6 shows the category breakdown. While our method does extremely well with 'significantly-more' pairs, it performs poorly in the 'almost-same' category.

Method	Weighted accuracy
Inter-human agreement	$92.80 \pm 0.68\%$
Our approach	$\mathbf{92.70} \pm 0.77\%$
Saliency detector	$89.26 \pm 1.20\%$
Center baseline	$86.07 \pm 1.08\%$
Scale baseline	$85.86 \pm 0.99\%$
Sharpness baseline	$88.52 \pm 1.13\%$

Table 5: Corpus-Level: Performance compared to baselines.

Pair category	Ours	S-Baseline	Improvement
significantly-more	96.35%	68.33%	28.02%
slightly-more	83.18%	71.82%	11.36%
almost-same	58.36%	69.93%	−11.57%

Table 6: Corpus-Level: Category-wise distribution of our predictions compared to Sharpness baseline.

Features	Image-Level	Corpus-Level
All	$92.72 \pm 0.93\%$	$92.70 \pm 0.77\%$
Without center	$91.25 \pm 0.95\%$	$92.41 \pm 0.71\%$
Without scale	$92.86 \pm 0.99\%$	$92.43 \pm 0.86\%$
Without sharpness	$92.22 \pm 1.10\%$	$91.52 \pm 1.31\%$
Only scale, center and sharpness	$89.53 \pm 1.13\%$	$90.54 \pm 1.81\%$

Table 7: Feature Ablation: Image-Level and Corpus-Level.

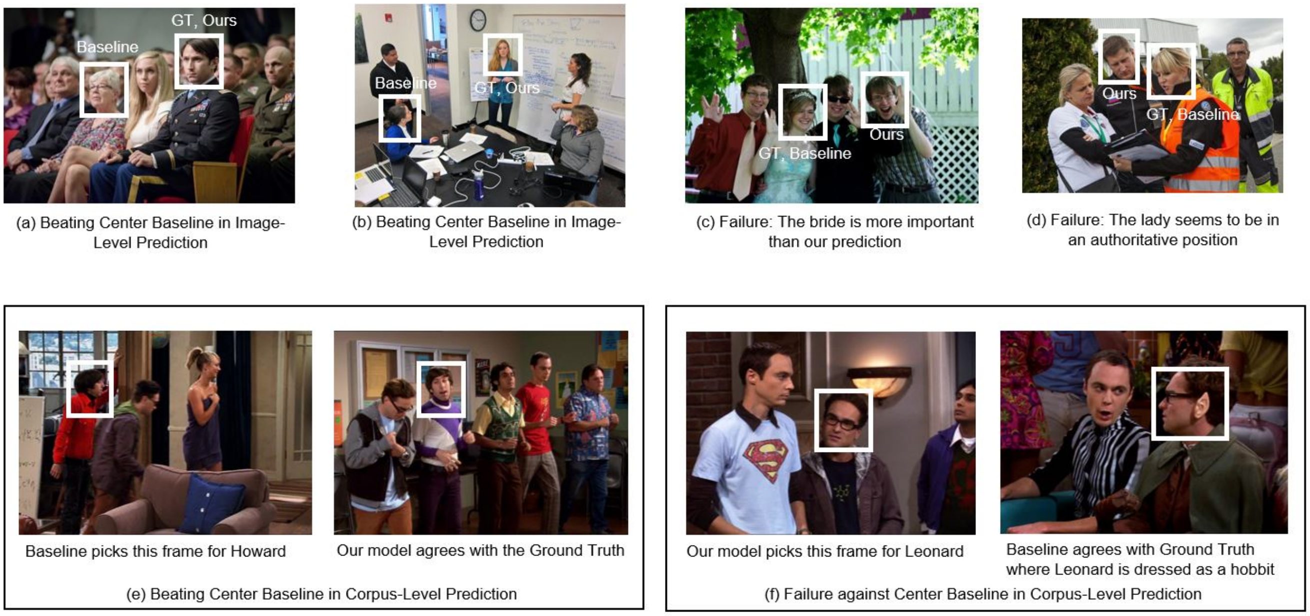
Fig. 4 also shows qualitative results for corpus experiments. Table 7 reports results from an ablation study, which shows the impact of the features on the final performance.

5. Importance vs Saliency

Now that we know we can effectively predict importance, it is worth investigating how importance compares with visual saliency. At a high level, saliency studies what draws a viewer's attention in an image. Eye-gaze tracking systems are often used to track human eye fixations and estimate pixel-level saliency maps for an image. Saliency is potentially different from importance because saliency is controlled by low-level human visual processing, while importance involves understanding more nuanced social and semantic cues. However, as concluded by [6], important objects stand out in an image and are typically salient.

We have already seen in Tables 3, 5 that saliency detectors perform worse than baselines in the image-level task and worse than our model in the corpus-level task respectively.

So how much does the salience of a face correlate with the importance of the person? We answer this question via the dataset collected by Jiang *et al.* [14] to study saliency in group photos and crowded scenes. The dataset contains eye



(a) Beating Center Baseline in Image-Level Prediction

(b) Beating Center Baseline in Image-Level Prediction

(c) Failure: The bride is more important than our prediction

(d) Failure: The lady seems to be in an authoritative position

Baseline picks this frame for Howard

Our model agrees with the Ground Truth

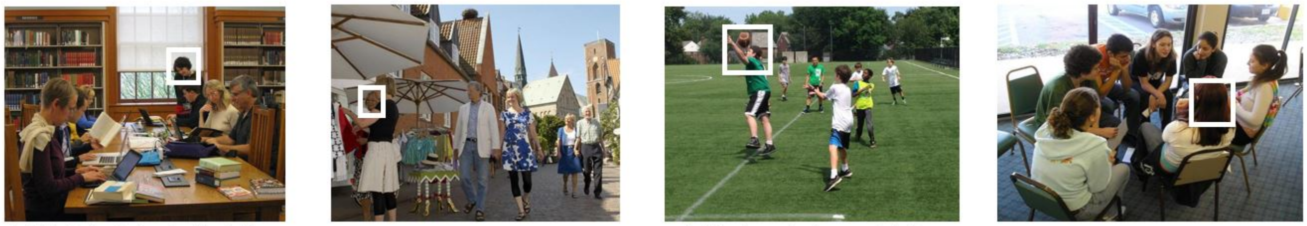
(e) Beating Center Baseline in Corpus-Level Prediction

Our model picks this frame for Leonard

Baseline agrees with Ground Truth where Leonard is dressed as a hobbit

(f) Failure against Center Baseline in Corpus-Level Prediction

Figure 4: Some results: (a)(b)(c)(d) for Image-Level prediction and (e)(f) for Corpus-Level prediction



(a) The lady who is standing is the most important but the least salient person – she appears to be a librarian or a supervisor

(b) The lady at the corner is the most salient but the least important person – the red colored garment could be the reason for more fixations points in that region

(c) The boy who is nearest to the football is both the most salient as well as the most important person

(d) This person is both the least salient and least important person in this image

Figure 5: Examples showing the relationship between visual saliency and person importance

Salience	Importance significantly more	 slightly more	 about same
significantly-more	38.33%	38.33%	23.33%
slightly-more	22.66%	32.81%	44.53%
about-same	03.82%	19.51%	76.67%

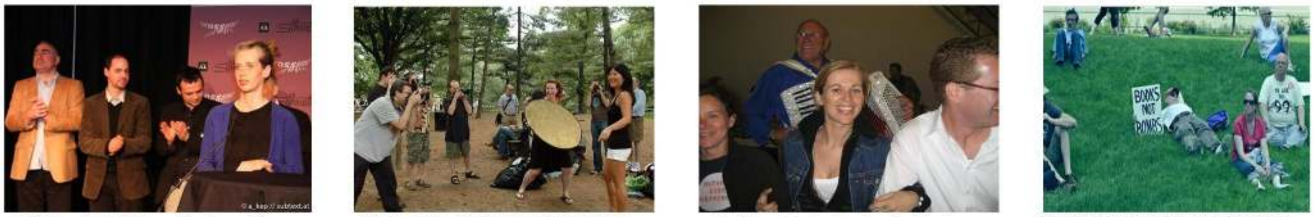
Table 8: Distribution of Importance pair categories among Salience pair categories

fixation annotations and face bounding boxes. For the purpose of this evaluation, we reduced the dataset to images with a minimum of 3 and maximum of 7 people, resulting in 103 images. In each image, the absolute salience of a face was calculated as as ratio of the fixation points in the face bounding-box to the total number of fixation points in all the face boxes in the image. This results in a ranking of people according to their saliency scores.

We then collected pairwise importance annotations for this dataset on Mechanical Turk using the same interface as used for the the Image-Level Importance dataset. Since this dataset is smaller, we annotated all possible face pairs (from the same image). Thus, we can extract a full ranking of individuals in each image based on their importance. Human judgement-based pairwise annotations are often inconsistent (*e.g.* $s_i > s_j$, $s_j > s_k$, and $s_k > s_i$). Thus, we used the Elo rating system to obtain a full ranking.

We measured the correlation between importance and saliency rankings using Kendall's Tau. The Kendall's Tau was 0.5256. The most salient face was also the most important person in 52.56% of the cases.

Fig. 5 shows qualitative examples of individuals who are judged by humans to be salient but not important, important but not salient, both salient and important, and neither. Table 8 shows the 'confusion matrix' of saliency *vs*. importance, broken down over the three strength categories. It can be seen that most face-pairs that are 'about-same' salient are also 'about-same' important whereas the other other two categories have less agreement – in a pair (p_i, p_j), p_i may be more salient than p_j but less important, and vice versa.



(a) **Ours & GT:** The white female in black shirt stands at the microphone. **Baseline**: The brunette male wearing a black shirt and jacket looks down.

(b) **Baseline & GT**: A smiling woman, in a black skirt is holding a round straw object. **Ours**: A man with glasses, wearing a light t-shirt is taking a photo.

(c) **All methods:** Lady with blonde hair smiling to camera in blue jean jacket.

(d) **GT**: A man is lying on the grass holding a sign. **Ours and Baseline**: A woman wearing sunglasses and a red shirt sits on the grass.

Figure 6: Qualitative results for the pruning descriptions experiment

6. Application: Improving Im2Text

We now show that importance estimation can improve im2text by producing more human-like image descriptions, as championed by the recent work of Vedantam *et al.* [27].

Sentence generation algorithms [15, 20] often approach the task by first predicting attributes, actions, and other relevant information for every person in an image. Then these predictions are combined to produce a description for the photo. In group photos or crowded scenes, such an algorithm would identify several people in the image, and may end up producing overly-lengthy rambling descriptions. If the relative importance of the people in the photo is known, the algorithm can focus on the most important people, and the rest can be either deemphasized or ignored as appropriate. How beneficial is importance prediction in such cases? This experiment addresses this question quantitatively.

Setup. Our test dataset for this experiment is a set of randomly selected 50 images from the Image-Level dataset. The training set comprises the remaining 150 images. Since the implementation for im2text methods was not available online at the time this work was done, we simulated them in the following way. First, we collected 1-sentence descriptions for every individual in the test set on Mechanical Turk. The annotation interface for these tasks asked Turkers to only describe the individual in question.

Prediction. We trained the importance model on the 150 training images and made predictions on the test set. We use the predicted importance to find the most important person in the image according to our approach. Similarly, we get the most important persons according to the center and random baselines. For each selection method, we choose the corresponding 1-sentence description. We then performed pair-wise forced-choice tests on Mechanical Turk with these descriptions, asking Turkers to evaluate which description was better, and found out the 'best' description per image.

Results. The importance methods were evaluated by how often their descriptions 'won' i.e., was ranked as the best description. The results in Table 9 show that reasoning about importance of people in an image helps significantly. Our approach outperformed the 'Random' baseline by 35%, which picks a human-written sentence about a random person in the image. An 'oracle' that picks the sentence corresponding to the most important person according to the ground-truth provides an upper bound (71.43%) on how well we can hope to do if we are describing an image with a single sentence about one person.

Method	Accuracy
Our approach	**57.14%**
Center	48.98%
Random	22.45%
Oracle	**71.43%**

Table 9: Importance prediction improves image descriptions: Each row reports the percentage of time the corresponding description was selected as the 'best' description.

7. Conclusions

To summarize, we proposed the task of automatically predicting the importance of individuals in group photographs, using a variety of features that capture the pose and arrangement of the people (but not their identity). We formulated two versions of this problem – (a) given a single image, ordering the people in it by relative importance, and (b) given a corpus of images for a person, ordering the images by importance of that person. We collected two importance datasets to evaluate our approach, and these will be broadly useful to others in the vision and multimedia communities.

Compared to previous work in visual saliency, the proposed person importance is correlated but not identical. Saliency is not the same as importance, and saliency predictors cannot be used in the place of importance predictors. People in photos may be salient but not important, important but not salient, both, and neither. Finally, we showed that our method can successfully predict the importance of people from purely visual cues, and incorporating predicted importance provides significant improvement in im2text.

The fact that our model performs close to the inter-human agreement suggests that a more challenging dataset should be collected. Compiling such a dataset, with richer attributes such as gender and age, and incorporating social relationship and popularity cues are the next steps in this line of work.

Acknowledgements. We thank Amir Sadovnik, Congcong Li, Makarand Tapaswi, and Martin Bauml for assisting us in various stages of our work. CSM and DB were partially supported by the Virginia Tech ICTAS JFC Award, the National Science Foundation under grants IIS-1353694 and IIS-1350553, the Army Research Office YIP Award W911NF-14-1-0180, and the Office of Naval Research grant N00014-14-1-0679. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government or any sponsor.

References

[1] CloudCV VIP demo. `http://cloudcv.org/vip/`. 5

[2] VIP project webpage. `https://computing.ece.vt.edu/~mclint/vip/`. 3, 4

[3] H. Agrawal, N. Chavali, C. Mathialagan, Y. Goyal, A. Alfadda, P. Banik., and D. Batra. CloudCV: Large-Scale Distributed Computer Vision as a Cloud Service. `http://cloudcv.org/`. 5

[4] A. Berg, T. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi. Understanding and predicting importance in images. In *CVPR*, 2012. 2

[5] A. Borji, D. N. Sihite, and L. Itti. Quantifying the relative influence of photographer bias and viewing strategy on scene viewing. *Journal of Vision*, 11(11):166, 2011. 5

[6] L. Elazary and L. Itti. Interesting objects are visually salient. *Journal of vision*, 8(3):3, 2008. 2, 6

[7] A. Gallagher and T. Chen. Understanding images of groups of people. In *CVPR*, 2009. 3

[8] A. C. Gallagher and T. Chen. Using group prior to identify people in consumer images. In *CVPR*, 2007.

[9] A. C. Gallagher and T. Chen. Finding rows of people in group images. In *IEEE International Conference on Multimedia and Expo*, pages 602–605, June 2009. 3

[10] J. Harel. A saliency implementation in matlab. http://www.klab.caltech.edu/ harel/share/gbvs.php. 5

[11] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in NIPS*, pages 545–552, 2006. 2

[12] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in NIPS*, 2006. 5

[13] S. J. Hwang and K. Grauman. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *International Journal of Computer Vision*, 100(2):134–153, Nov. 2012. 2

[14] M. Jiang, J. Xu, and Q. Zhao. Saliency in crowd. In *ECCV*. IEEE, 2014. 2, 6

[15] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on PAMI*, pages 2891–2903, 2013. 8

[16] D.-D. Le, S. Satoh, M. Houle, D. Phuoc, and T. Nguyen. Finding important people in large news video databases using multimodal and clustering analysis. In *2007 IEEE 23rd International Conference on Data Engineering Workshop*, pages 127–136, April 2007. 2

[17] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 2

[18] C. Li, A. C. Loui, and T. Chen. Towards aesthetics: A photo quality assessment and photo selection system. In *ACM Multimedia Conference*, 2010. 3

[19] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE Transactions on PAMI*, 33(2):353–367, 2011. 2

[20] A. Sadovnik, A. Gallagher, and T. Chen. Not everybody's special: Using neighbors in referring expressions with uncertain attributes. In *The V&L Net Workshop on Language for Vision, CVPR*, 2013. 3, 8

[21] H. Shu, A. C. Gallagher, H. Chen, and T. Chen. Face-graph matching for classifying groups of people. In *ICIP'13*, pages 2425–2429, 2013. 3

[22] SkyBiometry. `https://www.skybiometry.com/`. 3, 5

[23] M. Spain and P. Perona. Measuring and predicting object importance. *International Journal of Computer Vision*, 2010. 2

[24] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. "Knock! Knock! Who is it?" Probabilistic Person Identification in TV Series. In *IEEE Conference on CVPR*, Jun. 2012. 3

[25] B. W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14), 2007. 5

[26] P. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, and L. Itti. The impact of content-independent mechanisms on guiding attention. In *Proc. Vision Science Society Annual MeetingS*, May 2007. 5

[27] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726, 2014. 8

[28] X. Zhu and D. Ramanan. Face detection, pose estimation and landmark localization in the wild. In *CVPR*, 2012. 5