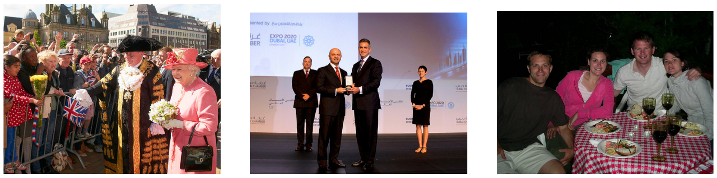
VIP: Finding Important People in Images

Clint Solomon Mathialagan[1], Andrew C. Gallagher[2], Dhruv Batra[1]
[1]Department of Electrical and Computer Engineering, Virginia Tech. [2]Google Inc.

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).



(a) Socially prominent. **(b)** Non-celebrities. **(c)** Equally important.

Figure 1: 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.

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 (Figure 2), 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 Figure 1a. 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.

Figure 1b 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. Figure 1c 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. Thus, the importance of a particular person in an image transcends the exact identity of the person, relying on social position, juxtaposition of people in the image, position relative to the camera, and probably many other factors

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 Figure 1a, 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.

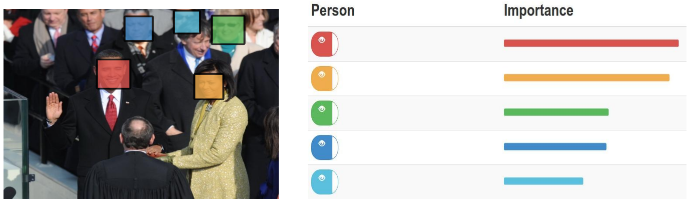


Figure 2: Goal: Predict the importance of individuals in group photographs (without assuming knowledge about their identities).

Approach. 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. An image-level dataset was created by mining images from Flickr.
- **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. Frames from the Big Bang Theory TV series were used to create the corpus-level dataset.

Learning. Importance is a subjective measure. We rely on the wisdom of the crowd to obtain ground truth for pair-wise importance comparisons in the datasets. We extract different visual features for persons from their face bounding boxes capturing the centerness, scale, sharpness, facial pose and occlusion. We model the task as a pair-wise regression problem and use support vector regression. The results show that we can easily beat the baselines and learn a reliable model.

Application. Is it useful to predict person importance? To gauge its practical utility, we explore combining with im2text for generating sentences for images with multiple people. Typically, sentence generation algorithms approach the task by first predicting attributes, actions, and other relevant information for every person. These inferences combine to produce a description for photo. For a group photo or crowded scene, the result is often an overly lengthy, rambling description. With our importance measure, the description can focus on the most important people, and the rest can be either deemphasized or ignored as appropriate.

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

Table 1: Better selection of sentences

First, we collected 1-sentence descriptions for every individual in a test set of 50 images using Mechanical Turk. We trained the importance model on 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. The methods were evaluated by how often their descriptions 'won' i.e., was preferred most often. The results in Table 1 show that reasoning about importance of people in an image helps significantly.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

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