EasyChair Preprint

№ 4391

Where's Wally: A Gigapixel Image Study for Face Recognition in Crowds

Cristiane Ferreira, Helio Pedrini, Wanderlay Alencar, William Ferreira, Thyago Peres Carvalho, Naiane Sousa and Fabrizzio Soares

October 13, 2020

Where's Wally: A Gigapixel Image Study for Face Recognition in Crowds

Cristiane B. R. Ferreira[1 [0000−0002−2180−4214]],
Helio Pedrini[2 [0000−0003−0125−630X]],
Wanderley de Souza Alencar[1 [0000−0002−3785−9527]],
William D. Ferreira[1 [0000−0003−2374−0653]],
Thyago Peres Carvalho[1 [0000−0002−1065−1961]],
Naiane Sousa[1 [0000−0001−7226−8928]], and
Fabrizzio Soares[3,1 [0000−0003−1598−1377]]

[1] Instituto de Informática, Universidade Federal de Goiás, Goiânia/GO, Brazil,
http://www.inf.ufg.br, cristianebrf@ufg.br,
wanderleyalencar@ufg.br, wferreira7@ufg.br,
thyagopcarvalho@gmail.com, naianesousa@discente.ufg.br
[2] Institute of Computing, University of Campinas, Campinas/SP, Brazil,
https://www.ic.unicamp.br/, helio@ic.unicamp.br,
[3] Department of Computer Science, Southern Oregon University, Ashland/OR, USA,
http://www.sou.edu, soaresf@sou.edu

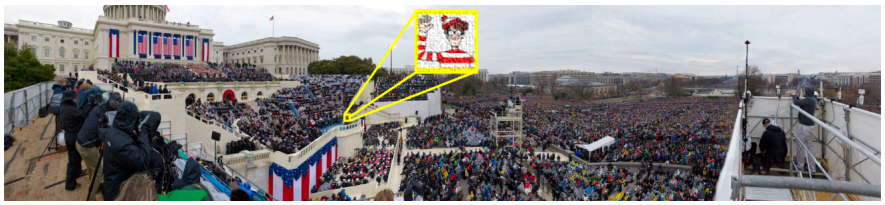


Fig. 1: Where's Wally[1] concept of our proposed study (Adapted from CNN).

Abstract. Several devices are capable of capturing images with a large number of people, including those of high resolution known as gigapixel images. These images can be helpful for studies and investigations, such as finding people in a crowd. Although they can provide more details, the task of identifying someone in the crowd is quite challenging and complex. In this paper, we aim to assist the work of a human observer with larger images with crowds by reducing the search space for several images to a ranking of ten images related to a specific person. Our model collects faces in a crowded gigapixel image and then searches for people using three different poses (front, right and left). We built a handcraft

[1] Where's Wally books are also known as Where's Waldo in North America.

dataset with 42 people to evaluate our method, achieving a recognition rate of 69% in the complete dataset. We highlight that, from the 31% "not found" among the top ten in the ranking, many of them are very close to this boundary and, in addition, 92% of non-matched are occluded by some accessory or another face. Experimental results showed great potential for our method to support a human observer to find people in the crowd, especially cluttered images, providing her/him with a reduced search space.

Keywords: Gigapixel image · Face detection · Face recognition · Crowd Visualization.

1 Introduction

A crowd of people is a common situation in which people agglomerate around an event for some reason. For instance, people usually gather in places for a common interest, such as outdoors, theaters, stadiums, shopping malls and airports. Port Authority of New York and New Jersey [8] reports that airports, such as the Hartsfield–Jackson Atlanta International Airport, had about 103,902,992 passengers only in 2017, which is approximately 284,665 per day or 11,861 per hour.

When there are hundreds or even thousands of people together, there may be situations where it is important and necessary to detect and identify people at the scene, such as a missing child or a crime suspect. In this way, equipment can capture images or videos of scenes typically populated by a large number of people and the identification of people on site is required for several reasons, including the occurrence of a crime and the search for a possible suspect, considering her/his facial identification. Identifying multiple faces in a crowd is not a simple task because it requires prior face separation and subsequent identification and classification of an already trained facial database.

Fig. 1 represents part of President Donald Trump's inauguration speech as the President of the United States. This image is known as a gigapixel image. Gigapixel images belong to a category that contains a very large amount of information, since their sizes can vary from 0.3 to 300 gigapixels or more. Generally, these images have hundreds of single pictures stitched together to create a huge, unique image.

Some studies show that the use of visual scenes that approximate the detail and complexity of natural scenes helps to understand the properties of a complex visual scene and influences its complete understanding, as shown by Clarke et al. [4]. Thus, human beings have great difficulties in finding patterns in images overloaded with information, such as images of crowds. Considering this, in our work we present a visual aid model for a human observer to work with crowded gigapixel images by reducing the search space for several images to a ranking of ten images related to a specific person. Our approach is able to separate faces in these images, as well as the search and recognition of people through a test dataset, considering a set of three different poses for each person to be found and

identified in the gigapixel image. Our model contributes to a better visualization and search space reduction for an observer's analysis.

Our text is organized as follows. Section 2 presents a short literature review related to gigapixel images and crowded gigapixel images. Section 3 describes the datasets used in our experiments to validate our method, described in Section 4. Section 5 reports and analyzes the experimental results. Finally, Section 6 concludes our work and presents directions for future work.

2 Literature Review

Some advances have been made in the field of gigapixel images. Yang et al. [11] used a deep convolutional network to discover responses of facial parts from arbitrary uncropped face images. Part detectors emerged within CNN trained to classify attributes from uncropped face images.

Bai et al. [2] presented a convolutional neural network architecture for face detection, where super-resolution and refinement network were used to generate real and sharp high-resolution images and a discriminator network was introduced to classify faces vs. non-faces. Furthermore, a loss was introduced to promote the discriminator network to distinguish the real/fake image and face/non-face simultaneously.

Zhang et al. [12] proposed a method that performs pedestrian counting on a gigapixel image. Pedestrian detection is performed using Exemplar Support Vector Machines running on a GPU-based architecture, and object-oriented histograms are used for the object characterization. Cao et al. [3] introduced a gigapixel crowd counting method. They used Dilated Convolution Neural Network and they performed the count on 3 different scales and weighted the results.

Wang et al. [10] presented a gigaPixel-level humAN-centric viDeo dAtaset, for large-scale, long-term and multi-object visual analysis, comparing human detection and tracking tasks and aiming to assist in the analysis of human behavior and interactions in large-scale real-world scenarios.

Ferreira et al. [5] presented a study and method to deal with pedestrian detection in gigapixel images by reducing their dimensions and associated computational effort. In their work, a multiresolution analysis was performed on gigapixel images to evaluate the required time processing and its impacts caused by the person detection algorithm.

Ferreira et al. [6] presented an impact analysis of the resolution reduction in the detection of people on gigapixel images. People detectors were trained with the INRIA and CALTECH data sets and results showed that, although gigapixel images provided a huge rate of false positives, the resolution reduction significantly decreased the number of bounding boxes and false positives, however, increased the rate of missing people.

3 Dataset Description

In our work, we used two datasets for combined experiments. The first one is a gigapixel image from Trump's inauguration, discussed in Subsection 3.1. The second one is a handcrafted dataset of different poses from different events of Trump's inauguration special guests, presented in Subsection 3.2.

3.1 Gigapixel Image Dataset

The gigapixel picture that we evaluated in our approach is named *The Inauguration of Donald Trump*, taken by CNN[2]. It was taken in Washington DC on January 20, 2017 at 3:00 PM GMT-2 and an estimated 300,000 to 600,000 people attended the public ceremony. Although, there is no information available on the equipment used to take the picture, in a StackExchange photography forum [9], there is a discussion speculating that it was taken with a Gigapan Epic Pro. However, more details on how CNN captured Trump's inauguration can be found in [1]. The image is stored in a cube face representation, where each cube face means a camera targeting direction. We extracted the image from `www.fanpic.co` with a Python script, via URL presented as follows:

```
http://europe.tiles.fanpic.co/749-2017-cnn/
  mres_{s}/l{l}/{v}/l{l}_{s}_{v}_{h}.jpg
```

where {s} is the cube face composed of f, b, u, d, l and r, which are front, back, up, down, left and right, respectively, {l} is the resolution level, where 1 is the lowest and 7 is the highest resolution, and {v} and {h} are the vertical and horizontal position of the camera, respectively, ranging from 1 to 125 at the highest resolution.

We downloaded the image of all cube faces at level 7, which is the highest resolution, and provides RGB tiles with 512×512 pixels and a total of 93,750 tiles. However, we used a small portion of the gigapixel image that corresponds to the surrounding location of the president's speech stage, where authorities sit, such as Supreme Court judges, Senators and Deputies, former presidents and first ladies, religious leaders, family and closer friends, and so on. We consider columns from 92 to 124 on the left cube face, columns from 00 to 30 on the front cube face and lines from 73 to 90 on both cube faces, therefore, resulting in a total of 1,152 tiles.

The tiles are designed to avoid spherical effect of the picture acquisition process and already stitched. However, there is no information about the software and algorithms used for that. In addition, we observed that, in several cases, flattening of the image occurs, which can bring some effects of radial distortion of people's faces and hinder the process of facial recognition.

[2] Available at `https://edition.cnn.com/interactive/2017/01/politics/trump-inauguration-gigapixel`. (Access was unstable when finishing this text).

It is also important to mention that, in the construction of this gigapixel image, several problems hinder the process of identifying and recognizing people, due to factors such as their position in the scene, for example. Several people appear from behind, others from the side and others with their faces hidden. Thus, there are several problems such as the variance of people's pose, occlusion, which are major challenges to be solved.

We created a strategy to go through the gigapixel image that will be detailed in Section 4 and we obtained a total of 432 face images with this approach after discarding a few non-faces detected images. Fig. 2 shows these faces and Table 1 shows some statistics for this face dataset.

Fig. 2: Gigapixel image samples to construct the gigapixel face dataset.

Table 1: Statistics for the gigapixel image dataset.

Images	Total	Face Position				Occlusion by Accessories	Occlusion by Other Faces	No Occlusion
		Frontal	Right	Left	Backface			
All	432	111	284	11	26	123	41	268
Selected	42	12	25	1	4	7	9	26

Table 1 groups faces based on two main criteria: face position and face visibility. For the first case, we observed four frequent positions in these images: front, right, left and back. In the second one, we observed images with accessory occlusion, images with face occlusion and no occlusion. Using part of the gigapixel image, our model found 432 faces. From this total and considering the first criterion, 111 faces were found in the front position, 283 faces turned to the right, 11 faces to the left and 26 with backface. Since there are many people present in the image who are not known authorities or public people, 42 people from this total were known (authorities, religious leaders, family members,

among others). Thus, considering this new total, we have 12 of them who appear with their faces in the front position, 25 with faces turned to the right, 1 with a face turned to the left and 4 with backface. When considering the second criterion, we are concerned with seeing if the faces are occluded. Most faces were not occluded (26) and 16 faces were occluded by accessories (7) or by another face (9).

Fig. 3 shows some of the problems found and which make it difficult to recognize the faces extracted from the gigapixel image. In Fig. 3(a), we can see a backface. In Fig. 3(b) to 3(f), we can observe that the faces are occluded by various accessories, such as hat, glasses, raincoat and hand on the face. In Fig. 3(g), (h) and (i), we have occlusion by other faces. In Fig. 3(j), we have a complex problem caused by gigapixel stitching distortion.

(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 3: Gigapixel dataset samples: (a) backface; (b) to (f) accessory occlusion; (g) to (i) occlusion by other faces and (j) gigapixel stitching distortion.

3.2 Handcrafted Dataset

This dataset was createdfrom images downloaded from people who were at Trump's inauguration, such as member of the U.S. Supreme Court, former presidents and first ladies, President Donald Trump and his family members, some religious celebrities, representatives and senators.

We use the images of the front, left and right poses of each individual's face, in different situations and dates, as shown in Fig. 4. In the handcrafted database, we have 42 different individuals, with 3 images of each, resulting in a total of 126 images.

We searched several websites for a complete list of all authorities, family, friends, religious leaders, Supreme court judges, speaker of the house representatives, among others, who were present at President Trump's inauguration. Since we have not found this complete document, we conducted searches on several sites, such as BBC, Wikipedia and U.S. Supreme Court, to find some names of people invited and who attended the event.

In addition, as President Trump's inauguration was not just attended by officials, we had difficulty finding a public dataset to collect images of everyone present at the ceremony. We found a dataset of images of members of the American Congress[3], but it consists of images only in the frontal pose of each member.

[3] `https://github.com/unitedstates/contact-congress`

However, in the image gigapixel, several people appear with their faces hidden or in a side pose due to the positioning of the equipment that captured the images. Thus, we chose to build our database considering the criteria of the three positions: front, right and left, as well as the issue of copyright of the images. The dataset construction was a very challenging task, as not all images available on the Internet are available for public use.

Left Frontal Right

Fig. 4: Example of Obama's faces in the handcrafted dataset.

4 Proposed Method

In our model, we developed two stages, both using the DLIB library for face detection and matching. The first is related to the detection of faces in the region of interest from President Trump's inauguration. The second is related to the use of 126 images of the dataset built with images of people known in three different poses and the attempt to locate these people on the faces detected in the gigapixel image. In Subsections 4.1 and 4.2, we describe details of our approach.

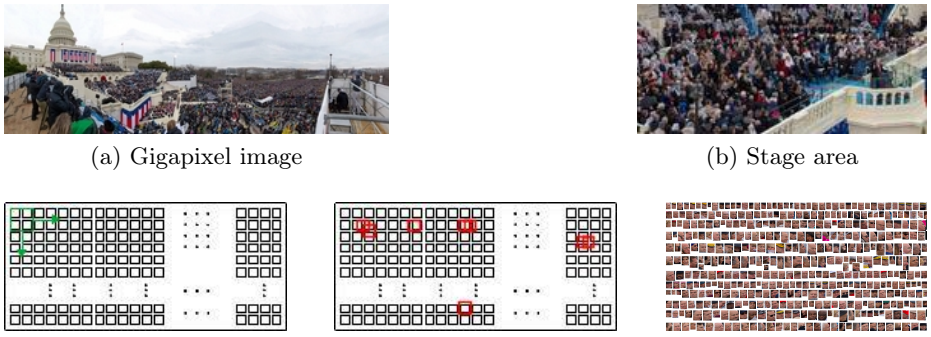
4.1 Face detection in Gigapixel Image

At this stage, we first select a small portion of the gigapixel image that corresponds to the location surrounding the president stage. Fig. 5(a) shows an entire gigapixel image and Fig. 5(b) shows the portion of the gigapixel image that we are considering in this work. Thus, we construct a model to scroll a sliding window through it, with 2 columns $\times$ 2 rows of tile images with 512$\times$512 RGB pixels each, resulting in a 1024$\times$1024 pixel sliding window, as seen in Fig. 5(c). This measurement was used considering the average face size in the gigapixel image.

We proceed to face detection by scanning the portion of the gigapixel image with the sliding window. As the stride adopted for the sliding window approach is 1 tile, the scanning process may find more bounding boxes than necessary, since many of them are overlapping. Although we could adopt strides larger than 1 tile, this could cause data loss, since we did not use any previous information about the faces in the images. Thus, in order to reduce unnecessary bounding

boxes and face redundancy, we performed a merging process by calculating the intersection over union (IoU) based on a non-maximum suppression criterion and we adopted a similar approach developed by Ferreira et al. [5,6]. If the IoU ratio is greater than a defined value, we merge the bounding boxes. In this study, we chose a threshold value of 0.9 for the fusion.

To find each bounding box that contains a face, we use the DLIB library in each window. This library uses the Histogram of Oriented Gradients (HOG) and considers the score of how many gradients points there are in each direction (up, right, etc). Then, the HOG method searches for the most similar to a known HOG pattern that was extracted from several other training faces. This approach uses the Face Landmark Estimation, proposed by Kazemi and Sullivan [7], whose main idea is to calculate 68 facial landmarks. Then, a machine learning algorithm is trained to find these 68 specific points on any face. This model trains a Deep Convectional Neural Network to generate an array of 128 measurements for each face. This embedding process is illustrated in Fig. 5(c) and 5(d). Fig 5(e) shows a sample of 432 face images detected in a small portion of the gigapixel image that corresponds to the location around the president's speech stage, where former presidents, former first ladies, senators, among other guests, were at the ceremony.



(a) Gigapixel image (b) Stage area

(c) Face detection (d) Face merging (e) Final dataset

Fig. 5: Face detection steps in the gigapixel image.

4.2 Matching between the Handcrafted Dataset and Gigapixel Face Images

In this step, we use a set of images of public people that include authorities and known people, who were present at the ceremony, considering 3 poses of each one (front, right and left). We built the image dataset shown in Section 3.2. An image sample is shown in Fig. 6a). Similar to the previous stage, we also performed a face detection step in the handcrafted dataset images through

DLIB, as shown in Fig. 6a), and extracted an array of 128 landmark features for each one. For face matching, we calculate the Euclidean distance (Eq. 1) between the landmark feature array of each image to be tested and all landmark feature arrays of face images extracted from a small portion of the gigapixel image considered, as shown in Fig. 6c). Finally, we build a distance ranking using Eq. 2 to be used as a measure to conclude whether a face was correctly recognized or not.

$$Distance_{i,j} = \sqrt{\left(\sum_{k=1}^{n} \left(fperson_i^p(k) - fgiga_j(k)\right)\right)^2}, \tag{1}$$

where $fperson_i$ and $fgiga_j$ are landmark feature sets of images i and j in the handcrafted and gigapixel datasets, respectively, p is a person pose (left, right, frontal), k is a landmark feature, n is the total amount of features ($n = 128$), and $Distance_{i,p,j}$ is the Euclidean distance between the landmark features.

$$Rank_i = \min\left(Distance_{i,j}\right), \tag{2}$$

where $Rank_i$ is the least distance of person i against gigapixel dataset images.

(a) Sample (b) Face detection (c) Face matching (d) Ranking

Fig. 6: Matching between the handcrafted dataset and gigapixel face images.

5 Experimental Results

We conducted our experiments to find people in the crowd considering only the president's speech surrounding area, which is reserved for the new president, vice president and their family and friends, former presidents, first ladies and former ones, Supreme Court judges, speaker of the House of Representatives, party leader, special authorities, invited guests, among others. For each person, we search for them using a frontal, left and right face pose. However, due to the camera's position, most people posed in right face. Since our goal is not to find an exact match, but to reduce an observer's effort, we considered the first 10 faces based on a measured distance ranking. We present our results in Table 2, which summarizes the results by total, by face pose and also grouped by occlusion.

As we can see from our results, 69% were found, while 31% were considered not found. Although the most expressive problem is with backface, they have only 4 people and, also, the backface does not provide enough landmark points

Table 2: Result by minimum ranking.

	Ranking	Ranking By Face Pose				Occlusion by	Occlusion by	No
	Total	Front	Right	Left	Back	Accessories	Other Faces	Occlusion
≤ 10	69%	83%	64%	100%	25%	57%	44%	92%
> 10	31%	17%	36%	0%	75%	43%	56%	8%

to be matched, but one person (25%) was found. The second most expressive result is the right face pose, in which a third is the person is not found. This is the largest group, because of the audience and camera's position. It is possible to observe that the most common reason for not finding someone's face is occlusion by accessory or by another face, since only 8% without occlusion are not found. Fig. 7 illustrates our experimental results.

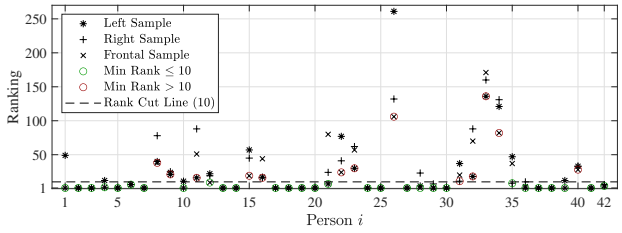


Fig. 7: Ranking of people found in our experiments.

We represent the pose of the person ranked bellow and above the top-10 (Rank Cut Line) as green and red circles, respectively. We can observe that, although there is group of people who did not rank among the first ten, many of them were very close. Moreover, many people have all poses ranked in close positions. It is important to note that, while $\approx 60\%$ of the faces in the gigapixel dataset are right poses, only $\approx 41\%$ of the matched faces are in fact right pose, whereas $\approx 33\%$ and $\approx 26\%$ matches are left and front poses, respectively. This shows that different poses are useful in the dataset, even if a specific pose is dominant in the crowd. Fig. 8 illustrates a histogram of the minimum ranking per person.

We can see that most of the images are at zero ranking and a minimum ranking of 10 is able to obtain an image cluster. If we increased the ranking to 12, 13, 14, 15 and so on, we could get one or two more images in each increment. However, that would increase the number of images to be inspected by an observer. We can also notice that some images are far from the defined boundary (10), for instance, there are images with a minimum ranking between 35 to 135.

Fig. 9 shows some interesting results achieved in our experiments. For example, in Fig. 9(a), a person is one of the faces ranked as first position for all

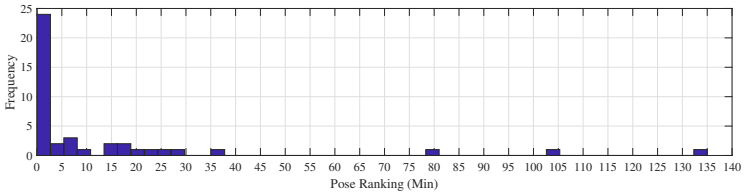
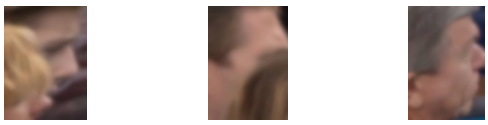


Fig. 8: Histogram of minimum ranking of person poses.

three poses, but a large part of his face is occluded by another face in the gigapixel image. Another more complex case, shown in Fig. 9(b), is a person who corresponds to only one of the three poses, even though his face was almost completely occluded by someone else. In addition, a challenging case due to backface is presented in Fig. 9(c), where the person falls in the first ten images in the face ranking, greatly reducing the amount of faces to be analyzed by an expert.



(a) (b) (c)

Fig. 9: Backface and occluded face samples recognized in our approach.

6 Conclusions

In this work, we present a model to assist a human observer in finding people in large images of crowds, reducing the search space for several images to a ranking of ten images related to a specific person. We conducted a face detection on a crowded gigapixel image and then a face recognition considering a set of three different poses for each person in a handcrafted dataset. Several challenges were addressed in the process, such as face occlusion.

Our purpose was not to find an exact match, but to reduce the effort of the observer or specialist, so that if the measured distance classifies a person among the first ten images, the person is considered as found. Our approach showed promising results, as we achieved recognition rates of 69% in total. When considering images with a ranking higher than 10 for all three poses, 92% of them have occlusions. The proposed method has great potential to help a human observer to find people in the crowd, especially in cluttered images, through a reduced search space.

Acknowledgment

Authors thank to CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil) for the partial support of this study.

References

1. Arthur, K.E.: How CNN Captured the Gigapixel Image of Trump's Inauguration (2017 (accessed August 3, 2020)), http://www.afd-techtalk.com/gigapixel-trump/
2. Bai, Y., Zhang, Y., Ding, M., Ghanem, B.: Finding Tiny Faces in the Wild With Generative Adversarial Network. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 21–30. Salt Lake City, UT, USA (Jun 2018)
3. Cao, Z., Yan, R., Huang, Y., Shi, Z.: Gigapixel-Level Image Crowd Counting using Csrnet. In: IEEE International Conference on Multimedia Expo Workshops. pp. 426–428 (2019)
4. Clarke, A.D.F., Elsner, M., Rohde, H.: Where's Wally: The Influence of Visual Salience on Referring Expression Generation. Frontiers in Psychology **4**, 329 (2013)
5. Ferreira, C., Soares, F., Pedrini, H., Bruce, N., Ferreira, W., Cruz Junior, G.: Multiresolution Analysis on Searching for People in Gigapixel Images. In: IEEE Canadian Conference on Electrical Computer Engineering. pp. 1–4. Quebec City, QC, Canada (May 2018)
6. Ferreira, C.B.R., Soares, F.A., Pedrini, H., Bruce, N.M., Ferreira, W., Junior, G.C.: A Study of Dimensionality Reduction Impact on an Approach to People Detection in Gigapixel Image. IEEE Canadian Conference on Electrical and Computer Engineering **43**(3), 122–128 (Aug 2020)
7. Kazemi, V., Sullivan, J.: One Millisecond Face Alignment with an Ensemble of Regression Trees. In: IEEE Conference on Computer Vision and Pattern Recognition (Jun 2014)
8. Port Authority of New York and New Jersey: 2017 Annual Airport Traffic Report (Apr 2018), retrieved: 2019-02-21
9. StackExchange: How did CNN take that Gigapixel Photo for Trump's Inauguration? (2017 (accessed August 3, 2020)), https://photo.stackexchange.com/questions/86489/how-did-cnn-take-that-gigapixel-photo-for-trumps-inauguration
10. Wang, X., Zhang, X., Zhu, Y., Guo, Y., Yuan, X., Xiang, L., Wang, Z., Ding, G., Brady, D., Dai, Q., Fang, L.: PANDA: A Gigapixel-Level Human-Centric Video Dataset. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3265–3275 (2020)
11. Yang, S., Luo, P., Loy, C.C., Tang, X.: From Facial Parts Responses to Face Detection: A Deep Learning Approach. In: IEEE International Conference on Computer Vision. pp. 3676–3684. Santiago, Chile (Dec 2015)
12. Zhang, S., Yu, S., Ma, Q., Shang, P., Gui, P., Wang, J., Feng, T.: Pedestrian Counting for a Large Scene Using a GigaPan Panorama and Exemplar-SVMs. In: 9th International Conference on Computational Intelligence and Security. pp. 229–235. IEEE (2013)