# MuseumVisitors: a dataset for pedestrian and group detection, gaze estimation and behavior understanding

# MuseumVisitors: a dataset for pedestrian and group detection, gaze estimation and behavior understanding

Federico Bartoli[1], Giuseppe Lisanti[1], Lorenzo Seidenari[1], Svebor Karaman[1,2] and Alberto Del Bimbo[1]

[1]{firstname.lastname}@unifi.it, University of Florence
[2]sk4089@columbia.edu, Columbia University

## Abstract

*In this paper we describe a new dataset, under construction, acquired inside the National Museum of Bargello in Florence. It was recorded with three IP cameras at a resolution of* $1280 \times 800$ *pixels and an average framerate of five frames per second. Sequences were recorded following two scenarios. The first scenario consists of visitors watching different artworks (individuals), while the second one consists of groups of visitors watching the same artworks (groups).*

*This dataset is specifically designed to support research on group detection, occlusion handling, tracking, re-identification and behavior analysis. In order to ease the annotation process we designed a user friendly web interface that allows to annotate: bounding boxes, occlusion area, body orientation and head gaze, group belonging, and artwork under observation. We provide a comparison with other existing datasets that have group and occlusion annotations. In order to assess the difficulties of this dataset we have also performed some tests exploiting seven representative state-of-the-art pedestrian detectors.*

## 1. Introduction

The interest for challenging and realistic datasets is raising in the computer vision and pattern recognition community. All recent major advancements in fundamental computer vision tasks have been driven by the release of large and challenging datasets. Public datasets are often associated with challenges in order to push researcher to develop algorithms and systems that advance the state-of-the-art. For tasks like object recognition, detection and segmentation the PASCAL VOC [19] datasets are a reference for the community. Recently the large scale taxonomy annotated dataset ImageNet [12] provided the sufficient amount of data to train large and deep neural networks [24]. Deep learning provided a new set of tools for object classification and detection researchers that could easily improve performance by simple transfer learning of models fitted on ImageNet [22, 7].

Large scale action recognition with trimmed and untrimmed videos have been recently proposed [30] with a challenge. This was the first attempt to release a large scale dataset, both in term of classes and samples. Moreover untrimmed sequences were released as test samples in 2014 in order to push research in action recognition towards detection, or temporal segmentation of actions of interest.

Recently the problem of group behavior understanding gained attention. Understanding group behavior is a challenging and sometimes ill defined problem. Some authors addressed the task of understanding collective behaviors like standing in a queue or crossing the road [1, 9]. Other authors have addressed the problem of person to person interaction, that can both happen in couples or groups. This kind of task stems from social studies and psychology. In some cases approaches are exploiting the social behavior to improve other, more basic, tasks like tracking [3, 29]. More recently researchers began to address the analysis of collective patterns. A typical task is the detection of F-formations [10]; F-formations are patterns that create when two or more individuals arrange spatially so that they have equal and direct access to the space between them. Therefore there exist multiple F-formation kinds depending both on the amount of participants and their spatial location and orientation. Being able to detect the presence and types of F-formations allows to roughly understand social behavior of observed people.

Person interaction is also mainly described by the so called attention, that is usually measured by recognizing where a person gaze is directed [5, 8]. Estimating people gaze can give a finer understanding of the relationship between a person and the environment.

At the core of user behavior understanding lays the com-

puter vision problem of pedestrian detection. Most of the measurement and descriptors proposed to understand collective behaviors and group formations need either gaze or people location. Moreover gaze can only be accurately estimated if the head is located correctly.

We believe that to allow researchers to explore the group behavior understanding extensively many heterogeneous annotations are needed. Gaze and people location in images are a must. Multi-camera setups are usual in real scenarios, therefore a modern dataset should include multiple partially overlapped views of a scene. The presence of groups will certainly generate occlusions among people so a desirable property of a dataset is also an annotation of occluded parts of each pedestrian. Finally environmental information such as accurate camera calibration and relevant object locations in a single real world reference may help analyzing not only the person-person interaction but also the person-object and person-scene interaction.

In this paper we are proposing MuseumVisitors a dataset for person and group behavior understanding on which tracking, detection and coarse gaze estimation can be evaluated. We recorded this dataset at National Museum of Bargello in Florence, Italy. We provide camera calibration and object locations. Moreover we developed a multi-user web-based annotation tool that will allow a continuous growth of the dataset in the upcoming years. Annotation of groups, identities and occluded parts are provided. The dataset has been recorded across different times of the day thus generating challenging sequences in term of lighting conditions. We thoroughly evaluate modern state-of-the-art pedestrian detection in different set-ups.

## 2. Existing dataset for group and occlusion detection

Person detection is widely studied in literature and many datasets have been publicly released, each one with different characteristics. However, there is a lack of datasets with group annotation, that can be used for example in group detection, tracking and behavior analysis. In this section we briefly review some currently available datasets that contain groups or occlusion annotations.

**Group detection** The CAVIAR dataset [6][1] was released in 2003 for behavior analysis purposes. It consists of two sets of experiments, each one composed by a set of video clips taken also from different cameras. These sequences were recorded acting out different scenarios of interest for different behaviors. In literature this datasets were mainly exploited for tracking purposes [2, 34]. It comes with groups annotations and it can be exploited for group detection, tracking or behavior analysis.

The Friends Meet (FM) dataset[2] was recently proposed in [3] specifically for group detection and tracking. It contains groups of people that evolve, appear and disappear spontaneously, and experience split and merge events. It is composed by 53 sequences, for a total of 16286 frames. The sequences are partitioned in a synthetic set without any complex object representation and dynamics, and a real dataset. The real dataset also contains bounding boxes annotations for each observed subject along with identities. We only consider the latter in Table 1. However, it was recorded from a single camera positioned far away from the observed plane, with a strong perspective and it can be really difficult to detect people on its frames since classic detectors are usually trained on frontal or lateral person images [11, 16].

The Images of Groups Dataset [21][3] is a collection of people images from Flickr obtained by performing three searches with some selected keywords. However, this dataset largely differs from the classic pedestrian detection datasets [11, 16] since it was mainly designed for social behavior analysis on single-shot images. In each image, the authors provide the group annotations along with the gender and the age category for each person.

**Occlusion detection** Recently a lot of techniques have been focusing on person detection with occlusions handling [26, 28, 31]. However, due to the lack of datasets with occlusion annotations it is always difficult to produce a quantitative measure of this phenomenon and compare with other methods.

The Daimler Pedestrian Detection Benchmark dataset [17][4] is a set of images captured from a vehicle-mounted calibrated stereo camera rig that is moving in an urban environment. It contains bounding boxes annotations for pedestrians and non-pedestrians in the scene. No additional annotation are provided about visible (or occluded) part of each pedestrian. However, the test set is split between non-occluded and partially-occluded.

The Caltech dataset [16][5] is composed of 250000 frames extracted from 10 hours of videos acquired from a vehicle driving through regular traffic in an urban environment. In this dataset individual pedestrians have been labeled as *Person* while large groups were delineated using a single bounding box and labeled as *People*. The authors also provided this dataset with the annotation for all the occluded pedestrians by labeling both the full extent of the pedestrian and the visible region. As described in the paper most of the pedestrians (70%) are occluded in at least one frame.

[1] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

[2] http://www.iit.it/it/datasets-and-code/datasets/fmdataset.html

[3] http://chenlab.ece.cornell.edu/people/Andy/ImagesOfGroups.html

[4] http://www.gavrila.net/Datasets/Daimler_Pedestrian_Benchmark_D/Daimler_Multi-Cue_Occluded_Ped/daimler_multi-cue_occluded_ped.html

[5] http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/

| Dataset | # cameras | # frames | # individuals | # pedestrians | density | Group | Person ID | Occlusion | Gaze | Video | Calibration |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MuseumVisitors | 3 | 4808 | 43 | 53389 | 11.1 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| CAVIAR Shop. Center [6] | 2 | 72515 | ∼237[8] | 179283 | 2.5 | ✓ | ✓ | | | ✓ | ✓ |
| Friends meet [3] | 1 | 10685 | – | – | – | ✓ | ✓ | | | ✓ | ✓ |
| Caltech [16] | 1 | 250000 | 2300 | ∼ 350000 | 1.4 | ✓ | ✓ | ✓ | | ✓ | |
| Daimler Ped. Det. [17] | | 21790 | | 88880 | 4.1 | | | ✓ | | ✓ | |
| CVC-05 Part. Occl. [25] | | 593 | | 2008 | 3.4 | | | ✓ | | | |
| CUHK occlusion [27] | | 1063 | | 10191 | 9.6 | | | ✓ | | ✓ | |

Table 1. Comparison between existing datasets for group and occlusion detection. Missing information are denoted with "–".

CVC-05 Partially Occluded Pedestrian dataset [25][6] is composed of 593 frames sampled from different sequences. It contains annotations only about the full bounding box of each pedestrian and does not provide any information about visible (or occluded) part of each target.

The CUHK occlusion dataset [27][7] for activity and crowded scenes analysis contains 1063 images divided in 10 clips with occluded pedestrians from other five datasets: Caltech [16], ETHZ [18], TUD-Brussels [32], INRIA [11], CAVIAR [6]. The authors also provided this dataset with both the full pedestrian bounding box and the visible (not occluded) bounding box part for each pedestrian along with a flag that separate occluded persons from non-occluded ones.

An overview about the datasets described in this section is given in Table 1. Here, for each dataset, we report some quantitative information: the number of cameras used (# cameras), the number of frames (# frames), the number of identities that can be used for tracking or re-identification (# individuals), the number of annotated bounding boxes (# pedestrians) and the number of annotated bounding boxes per frame (density). For each dataset we also report some properties, such as the availability of: group annotation (Group), person identity for each annotation (Person ID), occlusion information for each bounding box (Occlusion), Gaze information (Gaze) of body or head, video sequences or single-shot frames (Video) and calibration information (Calibration).

## 3. Design of the dataset

The dataset is extracted from video sequences recorded inside the National Museum of Bargello in Florence. The goal of this dataset is to provide an evaluation framework for all the components of a pipeline of computer vision tools aimed at understanding the behavior and interests of the visitors inside the museum. To be able to understand the visitors' behavior a computer vision system must first be able to robustly detect persons even when the visitors evolve in groups. Furthermore, visitor's face and body orientation together with the artworks positions can provide more precise clues to fully understand visitor interest.

[6] http://www.cvc.uab.es/adas/site/?q=node/7

[7] http://www.ee.cuhk.edu.hk/ xgwang/CUHK_pedestrian.html

[8] We determined the number of subjects from the available ground truth.

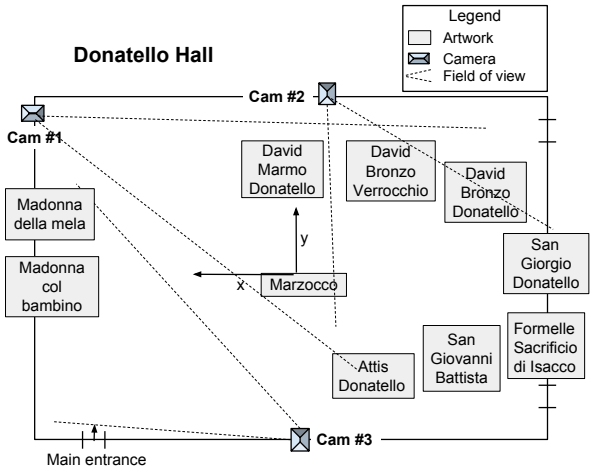



Figure 1. Scheme of the installation at the Bargello Museum with the 3 cameras positions and fields of view, artworks location and common ground plane axis.

In the following, we detail how the dataset was acquired and annotated.

### 3.1. Dataset acquisition

The installation at the Bargello Museum, depicted in Figure 1, makes use of 3 IP cameras connected to a local network through WiFi. Each camera video stream is acquired through a dedicated grabbing process at an average framerate of 5 frames per second. All cameras are calibrated to a common real world ground plane coordinates system, and the calibration information is released along the dataset. Furthermore, the real world coordinates of 10 artworks of interest inside the Donatello hall are recorded, enabling the dataset to be used for both behavior and interest analysis [23]. People filmed in the sequence were given very few instructions in order to avoid a choreographed behavior. Specifically each person or group was asked to visit a subset of the artworks with no specific order.

### 3.2. Annotation protocol

The dataset is annotated with different information about each person. First of all a bounding box enclosing each person is defined. If a person is partially occluded, a secondary bounding box annotation corresponding only to the visible part of the person is defined, see Figure 2(a). Each person is

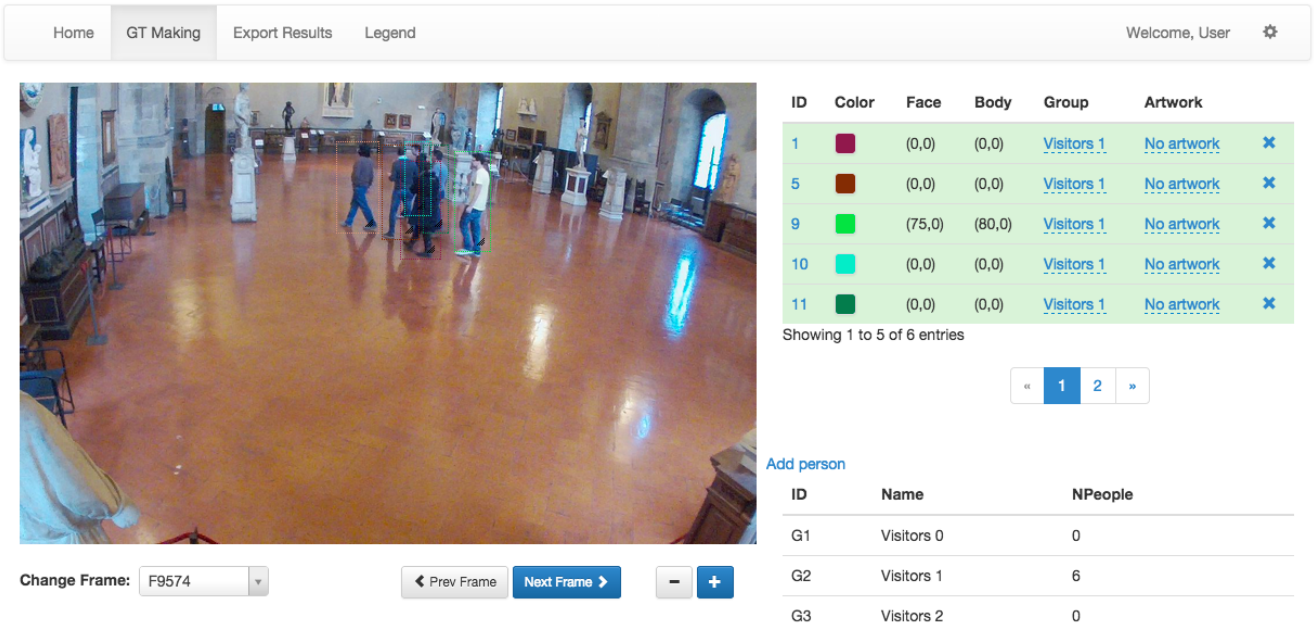


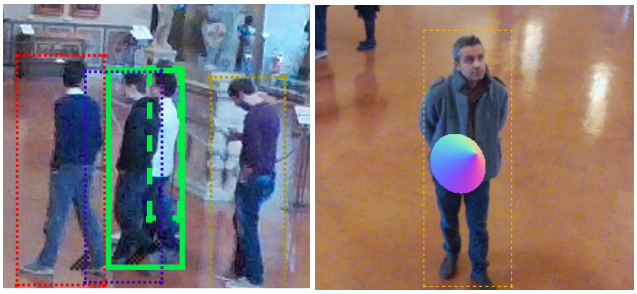

Figure 3. Annotation tool web interface.

associated with a single identifier on all frames of all cameras. If a person is part of a group, it is associated with the group identifier that is also shared on all frames of all cameras. Finally, the body orientation and gaze are also annotated according to a quantization of 5 degrees as shown in Figure 2(b).

### 3.3. Ground truth annotation tool

In order to ease the annotation process we designed a user friendly web interface shown in Figure 3. A great advantage of implementing the tool as a web application is the possibility of multi-user concurrent annotation.

As it is possible to observe in Figure 3, on the top we have a menu bar with different options: *GTmaking*, *Export results* and *Legend*. By choosing *GT making* the annotation tool asks for the user for its username and allows to chose the camera and frame to annotate, if none is specified the annotation process will start from the latest frame annotated by the user.



(a) (b)

Figure 2. *(a)* The solid green rectangle represent the bounding box selected for the annotation while the green dashed rectangle represent the visible (not occluded) area annotated by the user; *(b)* The cone visualizes the annotation of the gaze provided by the user.

On the *left-top* part of the interface, we show the chosen frame along with the already annotated bounding boxes. By selecting one of the bounding boxes the dashed rectangle become solid the user is able to move and resize the bounding box or to specify different information about that annotation, such as: the visible area (occlusion), the direction of the body and the gaze. A new bounding box can be added by clicking the "Add person" button or by pressing "+" on the keyboard. On the *left-bottom* part of the interface, we put some functional buttons that allow to navigate through the frames and zoom-in or out on the image (zooming is also possible by mouse scrolling). In the *right-top* part of the interface we put one table summarizing the information about each individual, like the person identifier (ID), the color of the bounding box, the gaze direction (Face), the body direction (Body), the group of which the selected user is part of (Group) and if it is standing by a particular artwork or not (Artwork). Finally, in the *right-bottom* part of the interface we put, instead, a table summarizing the groups information, like the identifier of the group (ID), the name of the group (Name) and the number of persons that are part of the group (NPeople).

In order to make this tool intuitive and ease the annotation process we defined a series of keyboard shortcuts to speed-up the process. These shortcuts are summarized in the *Legend* section of the annotation tool. Moreover, once

(a) *Camera 1 - Individuals* (b) *Camera 2 - Individuals* (c) *Camera 3 - Individuals*

(d) *Camera 1 - Groups* (e) *Camera 2 - Groups* (f) *Camera 3 - Groups*

Figure 4. Sample frames showing the different cameras and scenarios of the MuseumVisitors dataset.

| | Pedestrians height | | |
|---|---|---|---|
| Camera | Min | Max | Avg |
| 1 | 30 | 498 | 137 |
| 2 | 79 | 442 | 159 |
| 3 | 96 | 423 | 153 |

Table 2. Statistics about the pedestrians height (in pixels) in each camera of the dataset.

a frame is annotated, the successive frame will have the same bounding boxes as a starting point for the new annotations, in order to overcome the necessity of re-defining from scratch every person annotation at every frame.

## 4. Experiments

We performed a series of experiments to assess the difficulty of the MuseumVisitors dataset. Tests were conducted considering the frames extracted from the three cameras in the Donatello Hall, under two scenarios: individual and groups. The first scenario shows visitors watching different artworks, while the second one shows groups of visitors watching the same artworks. Figure 4 shows some sample frames for the different cameras and scenarios of the MuseumVisitors dataset. In Table 2 we report the minimum, maximum and average heights in pixels of all annotated visitors for each camera of the dataset.

We evaluated the proposed dataset with seven representative state-of-the-art pedestrian detectors [11, 20, 15, 14, 4, 13, 33]. One of the first successful approach to object detection has been proposed by Dalal *et al*. [11], designing a feature based on histograms of oriented gradient (HOG) and linear SVM. This detector has issues with deformable objects using a single holistic template, therefore Felzenszwab *et al*. [20] proposed a mixtures of part-based deformable models (DPM) in order to improve the detection of the targets in presence of occlusion and crowd in the scene. Recently several classifiers based on Haar-like features computed on multiple channels and soft-cascades have been proposed [15, 14, 4, 13]. This recent line of work obtain state of the art performances on challenging datasets [16] and lean towards efficiency. In [15] the Haar-like feature are computed, in an efficient way, over multiple channels by the Integral Channel Feature structure (ChnFtrs), which allows to reduce the computational effort without loss of accuracy in the detection process. In [14] (FPDW) the full pyramid features is approximated by interpolation at nearby scales, requiring only the exact computation of the feature in the middle-levels of each octave of the pyramid. In [4] the authors propose the VeryFast detector composed of multiple classifiers, each one trained for a specific octave of the pyramid. This in combination with the features approximation of [14] moves the feature extraction complexity from test time to training time. In [13] the authors proposed the Aggregate Channel Feature (ACF) extending the work in [14] with a variant of integral channel features to compute the pyramid features efficiently. The ACF detector was recently extended in [33] by applying a set of decorrelating filters per channel (ACF-LDCF).

For each detector we specify if it was trained on the INRIA pedestrian dataset [11] (I), on the Caltech pedestrian dataset [16] (C), or both of them (I+C).

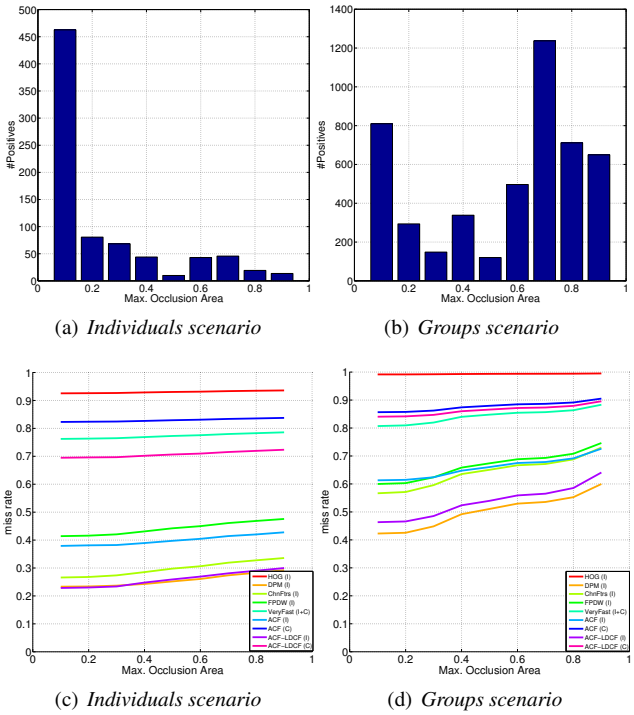
We performed an experiment to evaluate how occlusion influences the performance of tested detectors. As it can be observed from Figure 5(a) for the individual scenario most of the annotated bounding boxes have less than $10\%$ of occlusion level. This can be also noticed from Figure 5(c) where the performance of each detector does not vary too much as the occlusion percentage increases. On the contrary, for the groups scenario, the number of bounding boxes per occlusion level varies consistently (see Figure 5(b)) and this can be noticed from the fact that the performance of tested detectors decreases according to the occlusion level percentage, see Figure 5(d). With this result in mind and also inspired by [16] we designed a Reasonable experimental setting restricting pedestrian bounding boxes to be wider than 50 pixels and with less than 30% of occlusion. This restricted dataset setting removes objects that are very hard to detect either because their size is too small or because the occlusion does not provide enough evidence to the trained classifiers.

In Table 3 we report the accuracy obtained from the tested pedestrian detectors on the proposed dataset. Performances are summarized using the miss rate (MR) at $10^{-1}$ false positive per image (FPPI) for the three cameras. We report separately MR@$10^{-1}$ on the Full scenarios Individuals (Ind.), Groups (Group), and their respective reasonable versions (Reas.). We obtain different results for the three cameras due to the difference in terms of scales and locations of the visitors in the scene. For the individuals scenario the best performance are obtained with the DPM detector in the camera 3 ($32\%$), while the detector ChnFtrs is the best in the other cameras, with a MR of $67\%$ and $51\%$ respectively. For the groups scenario the best performance is obtained by the FPDW detector for both camera 1 ($89\%$) and camera 2 ($32\%$), while for the camera 3 the DPM detector reach the lower miss rate ($60\%$).

If we consider the reasonable setup all detectors have an higher accuracy drastically reducing all the Miss Rates on every camera. In particular, for the case of individuals the best result is obtained in the camera 3 with the ACF-LDCF(I) detector ($23\%$), while the best performer for camera 2 is the FPDW detector ($29\%$), and the ChnFtrs detector for the camera 1 ($57\%$).

The ROC curves of all the tested methods are reported in Figure 6 separately for individuals and groups and for each camera considering the Full scenario. While in Figure 7 we report the ROC curves separately for individuals and groups and for each camera considering the Reasonable scenario.

In general there is not a single pedestrian detector which obtains the best results in all sequences. This is due to the different complexities in each scenario that must be addressed by a single pedestrian strategy. This fact shows that the proposed dataset contains many challenges for pedestrian detection stemming from occlusion, lighting and scale changes that are inherent in a real world scenario.




(a) *Individuals scenario* (b) *Groups scenario*

(c) *Individuals scenario* (d) *Groups scenario*

Figure 5. Number of bounding boxes for both the individuals (a) and groups (b) scenarios for all the cameras varying the occlusion area. Average miss rate @$10^{-1}$ averaged over the three cameras for both individuals (c) and groups (d).

## 5. Discussion and Conclusion

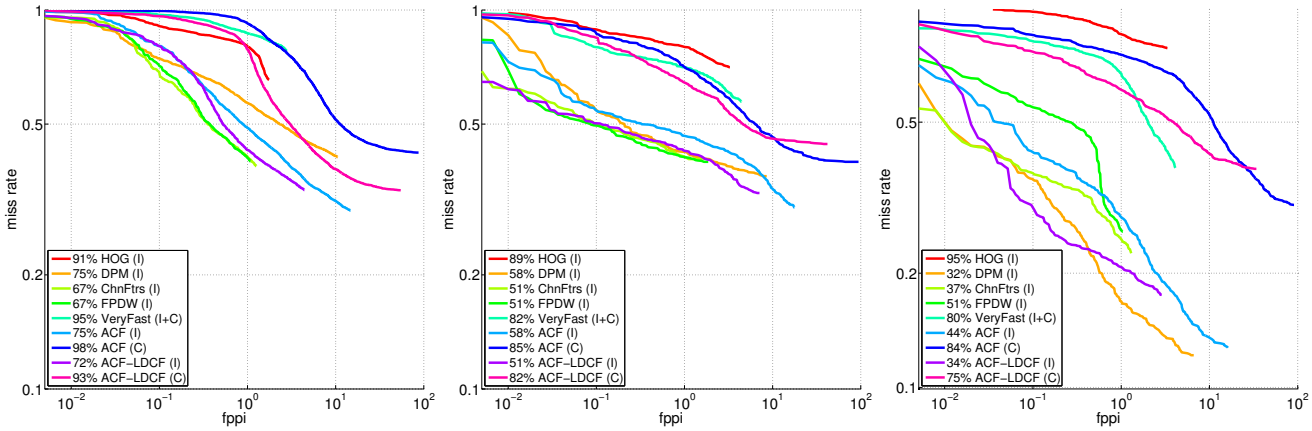
In this paper we presented a new dataset to serve many purposes and with unique characteristics. The MuseumVisitors dataset is a perfect testing ground for core computer vision techniques used as prerequisites for group behavior understanding such as: pedestrian detection under occlusion, group detection, re-identification, tracking and gaze estimation. We provide a level of detail in the annotation that lacks in many of the recent surveillance datasets. We propose several subsets of the dataset based on different scenarios such as: groups or individuals and full or reasonable scenarios; all of these scenarios are available for the three views.

The three views being calibrated on a single world coordinates reference system it is possible to combine the information gathered from multiple cameras at no cost. Furthermore, the real word coordinates of the artworks in the observed museum room are also given with the dataset. Hence, people behavior can be analysed in terms of relationship between individuals and relationships between individuals and the objects in the scene.

The dataset footage has been captured from a real system installed in a major Museum of the city of Florence pro-

| Detector | Camera 1 | | | | Camera 2 | | | | Camera 3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ind. | Ind. Reas. | Groups | Groups Reas. | Ind. | Ind. Reas. | Groups | Groups Reas. | Ind. | Ind. Reas. | Groups | Groups Reas. |
| HOG (I) | 91 | 88 | 99 | 96 | 89 | 80 | 98 | 97 | 95 | 93 | 100 | 99 |
| DPM (I) | 75 | 69 | **89** | 77 | 58 | 37 | 52 | 41 | **32** | 24 | **60** | **45** |
| ChnFtrs (I) | **67** | **57** | 90 | 74 | **51** | **29** | 42 | 32 | 37 | 27 | 73 | 60 |
| FPDW (I) | **67** | 58 | **89** | **72** | **51** | **29** | **41** | **31** | 51 | 42 | 75 | 62 |
| VeryFast (I+C) | 95 | 94 | 98 | 94 | 82 | 72 | 88 | 82 | 80 | 76 | 88 | 82 |
| ACF (I) | 75 | 70 | 91 | 80 | 58 | 48 | 55 | 47 | 44 | 38 | 73 | 62 |
| ACF (C) | 98 | 93 | 100 | 96 | 85 | 79 | 90 | 88 | 84 | 82 | 91 | 86 |
| ACF-LDCF (I) | 72 | 65 | **89** | 75 | **51** | 36 | 47 | 38 | 34 | **23** | 64 | 49 |
| ACF-LDCF (C) | 93 | 91 | 98 | 96 | 82 | 74 | 75 | 70 | 75 | 70 | 90 | 85 |

Table 3. Miss Rates @$10^{-1}$ False Positive per Image (fppi) of leading pedestrian detectors on the MuseumVisitors dataset. For each camera we evaluated the individuals (Ind.) and groups (Groups) scenarios, considering also the reasonable ground truth (Reas.). In bold we report the best results for each scenario.




(a) *Camera 1 - Individuals* (b) *Camera 2 - Individuals* (c) *Camera 3 - Individuals*

(d) *Camera 1 - Groups* (e) *Camera 2 - Groups* (f) *Camera 3 - Groups*

Figure 6. Evaluation results for the three cameras, on individuals and groups scenarios over all the dataset.

viding challenging crowding and lighting conditions. This setup will allow us to gather more sequences in the future and release subsequent, enlarged, versions of the MuseumVisitors dataset.

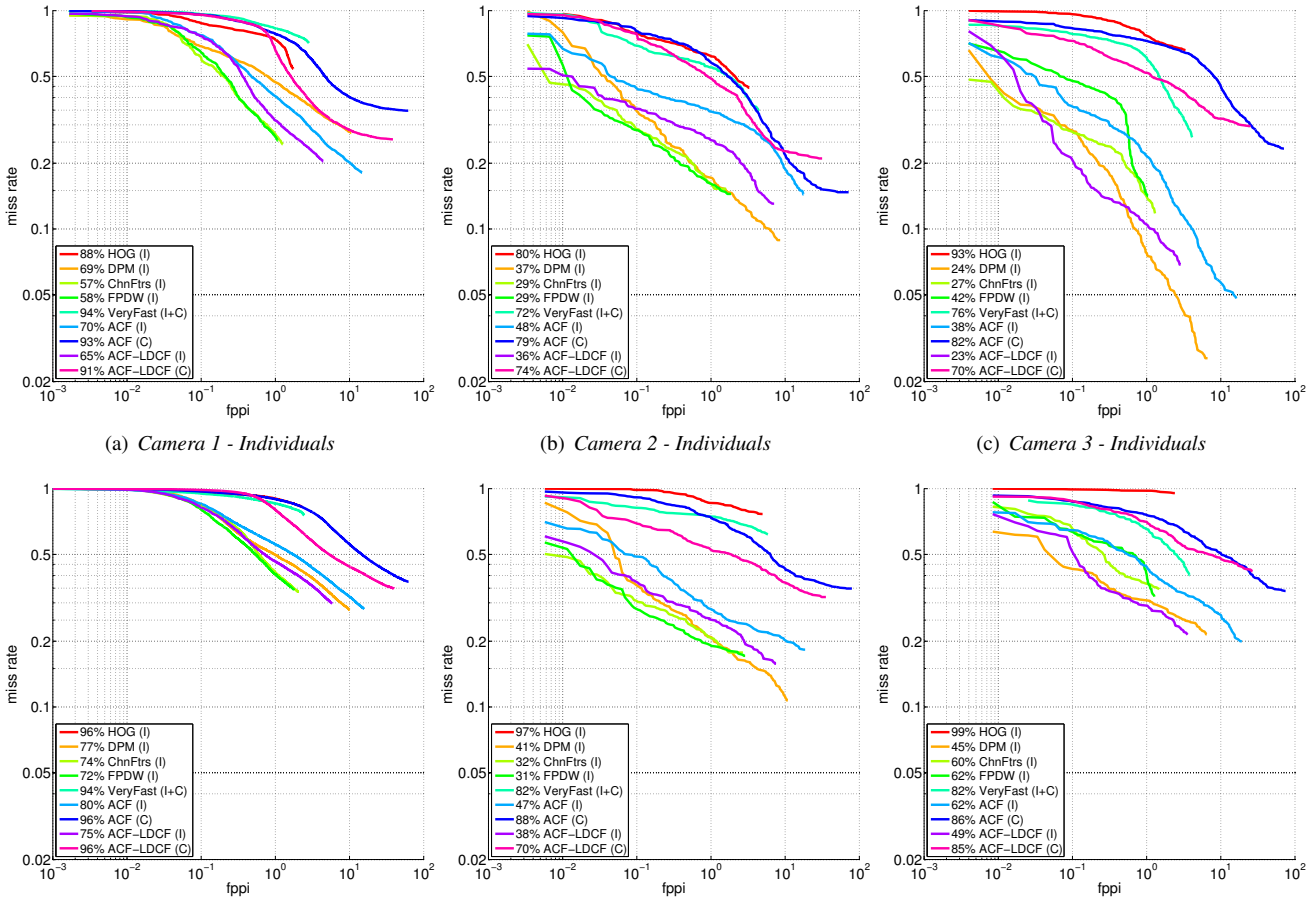
Having developed a user friendly, multi-user, web based annotation tool we are able to do a continuous annotation of the footage we have acquired and we have yet to release.

The annotation tool and the dataset will be publicly released with the data to generate our experimental results in order to ease future comparison with forthcoming methods.

## References


[1] M. Amer, P. Lei, and S. Todorovic. Hirf: Hierarchical random field for collective activity recognition in videos. In *Proc of ECCV*, 2014. 1

[2] S.-H. Bae and K.-J. Yoon. Robust online multi-object track-




(a) *Camera 1 - Individuals* (b) *Camera 2 - Individuals* (c) *Camera 3 - Individuals*

(d) *Camera 1 - Groups* (e) *Camera 2 - Groups* (f) *Camera 3 - Groups*

Figure 7. Evaluation results for the three cameras, on individuals and groups scenarios only over the reasonable annotations.


ing based on tracklet confidence and online discriminative appearance learning. In *Proc. of CVPR*, June 2014. 2

[3] L. Bazzani, V. Murino, and M. Cristani. Decentralized particle filter for joint individual-group tracking. In *Proc. of CVPR*, 2012. 1, 2, 3

[4] R. Benenson, M. Mathias, R. Timofte, and L. J. V. Gool. Pedestrian detection at 100 frames per second. In *Proc. of CVPR*, 2012. 5

[5] B. Benfold and I. Reid. Unsupervised learning of a scene-specific coarse gaze estimator. In *Proc. of ICCV*, 2011. 1

[6] CAVIAR. Test case scenarios. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/. 2, 3

[7] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proc. of BMVC*, 2014. 1

[8] C. Chen and J. Odobez. We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video. In *Proc. of CVPR*, June 2012. 1

[9] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *Proc. of ECCV*, 2012. 1

[10] M. Cristani, L. Bazzani, G. Pagetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino. Social interaction discovery by statistical analysis of f-formations. In *Proc. of BMVC*, 2011. 1

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of CVPR*, 2005. 2, 3, 5

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. of CVPR*, 2009. 1

[13] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *PAMI*, 2014. 5

[14] P. Dollár, S. Belongie, and P. Perona. The fastest pedestrian detector in the west. In *Proc. of BMVC*, 2010. 5

[15] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *Proc. of BMVC*, 2009. 5

[16] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 34(4):743–761, April 2012. 2, 3, 5, 6

[17] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *Proc. of CVPR*, 2010. 2, 3

[18] A. Ess, B. Leibe, and L. Van Gool. Depth and appearance for mobile scene analysis. In *Proc. of ICCV*, October 2007. 3

[19] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, June 2010. 1

[20] P. Felzenszwalb, R. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *Proc. of CVPR*, 2010. 5

[21] A. Gallagher and T. Chen. Understanding images of groups of people. In *Proc. of CVPR*, 2009. 2

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. of CVPR*, 2014. 1

[23] S. Karaman, A. D. Bagdanov, L. Landucci, G. D'Amico, A. Ferracani, D. Pezzatini, and A. Del Bimbo. Personalized multimedia content delivery on an interactive table by passive observation of museum visitors. *Multimedia Tools and Applications*, pages 1–25, 2014. 3

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. of NIPS*. 2012. 1

[25] J. Marin, D. Vazquez, A. Lopez, J. Amores, and L. Kuncheva. Occlusion handling via random subspace classifiers for human detection. *IEEE Transactions on Cybernetics*, 44(3):342–354, 2014. 3

[26] M. Mathias, R. Benenson, R. Timofte, and L. Van Gool. Handling occlusions with franken-classifiers. In *Proc. of ICCV*, 2013. 2

[27] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *Proc. of CVPR*, 2012. 3

[28] W. Ouyang and X. Wang. Single-pedestrian detection aided by multi-pedestrian detection. In *Proc. of CVPR*, 2013. 2

[29] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proc. of ICCV*, 2009. 1

[30] K. Soomro, A. Roshan Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. In *Proc. of CRCV-TR-12-01*, 2012. 1

[31] P. Wohlhart, M. Donoser, P. M. Roth, and H. Bischof. Detecting partially occluded objects with an implicit shape model random field. In *Proc. of ACCV*, 2012. 2

[32] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. In *Proc. of CVPR*, 2009. 3

[33] J. H. H. Woonhyun Nam, Piotr Dollár. Local decorrelation for improved pedestrian detection. In *Proc. of NIPS*, 2014. 5

[34] B. Yang and R. Nevatia. Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In *Proc. of CVPR*, 2012. 2