UNIVERSITÀ
DEGLI STUDI
FIRENZE

Università degli Studi di Firenze
Dipartimento di Ingegneria dell'Informazione (DINFO)
Corso di Dottorato in Ingegneria dell'Informazione
Curriculum: Informatica (ssd ing/inf-05)

User interest profiling by real time person detection and coarse gaze estimation

Candidate
Federico Bartoli

Supervisors
Prof. Alberto Del Bimbo

Dr. Giuseppe Lisanti, Ph.D

Dr. Federico Pernici, Ph.D

PhD Coordinator
Prof. Luigi Chisci

ciclo XXIX, 2013-2016

Università degli Studi di Firenze, Dipartimento di Ingegneria
dell'Informazione (DINFO).

Alla mia Famiglia.

Acknowledgments

I would like to acknowledge the efforts and input of my supervisor, Prof. Alberto Del Bimbo, and all my colleagues of the Media Integration and Communication Center (MICC) who were of great help during my research. A special acknowledgement goes to Giuseppe Lisanti, Svebor Karaman and Lorenzo Seidenari who directly collaborated on the main parts of my research work.

Contents

Chapter 1

Introduction

Understanding the focus of attention is a challenging computer vision task with many valuable and interesting applications. Attention may be directed towards other people or objects in the scene and both these cases represent strong cues in understanding people behavior. For the first case, usually referred as social signals and/or group behavior analysis, a reliable prediction on who is looking at whom is the main cue to seek. Group behavior is often defined in terms of spatial disposition and orientation of persons (people formations). However, body orientation estimation without gaze information may often lead to ambiguous predictions. Understanding instead what objects are looked at and for how long is also of great interest for retail companies that may want to obtain a large dataset of customer behavior. This is often solved by tracking all the persons in the scene and consequently generating heat images, registered with the shop maps, that indicate customer persistence. Although, even if the scene strongly constrains people position, such as in a supermarket aisles, there is a lot of ambiguity if we consider just the position. If we are willing to detect which products draw people attention in a shop, gaze estimation is the only option.
A slightly different but complementary task is profiling the interests of a single person in a given environment. In this case, instead of accumulating a global statistic from all persons behavior, a single profile is sought. In particular, given a set of person detections, the goal is to build identities and the corresponding interest profiles. Identity building is a problem similar to clustering and is usually solved exploiting person re-identification algorithms [62]. Once a certain amount of detections of a single individual are

connected an interest profile can be built. In this situation a higher precision is required since the amount of samples are scarcer.
Passive profiling finds several interesting applications in the cultural heritage scenario [53]. For example, user profiling can help solving many issues Museums struggle to cope with; like personalizing content for visitors. Personalization should both increase engagement and satisfaction creating a dedicated view of museum collections and suggesting novel cultural paths to explore. Moreover a recommender system may, also building from previously watched people behaviors, help in planning further tours towards different cultural venues, places of interest or museums.
To this end both person gaze and position in the scene are very crucial tasks for a correct attention estimation; how far an object is from the person could not be a sufficient hint. Gaze is usually inferred through head/body pose estimation which requires fast and accurate methods to detect the person.
Person detection is important in many video surveillance tasks of computer vision, like tracking [52, 80], person re-identification [59, 60] and human behavior analysis [28,58]. However, designing a generic pedestrian detector that works reliably and efficiently on different scenes remains a challenging task. Difficulties arise mainly because of changes in the camera viewpoint, different illumination conditions as well as the distinct backgrounds that characterize each scenario. Significant advances on person detection have been proposed in many research works in the last decade, as surveyed in [13, 38, 97]. These solutions focus on improving one or more stages of a person detector pipeline, as shown in Figure 1.1, which usually includes: pyramidal representation of the imaged frame to account for the different scales at which a target may appear; mapping of image content into a feature space to improve the capability to discriminate within the scene; content classification in the detection windows; and selection of the windows that most likely contain a target. Typically, a detection window of fixed size is slided over a predefined grid with a fixed sampling rate until the frame is fully scanned. Since persons can appear at different scales, the process is repeated on each rescaled image of the pyramid. However, in most of the cases, both sliding the window over the entire image at different scales and feature extraction and classification at each window have shown to be prohibitively costly to run in real-time [33, 37]. Solutions to expedite the different stages of the pipeline have addressed feature computation with new efficient and effective feature extraction methods [14, 34, 35, 98], or introduced strategies to reduce the

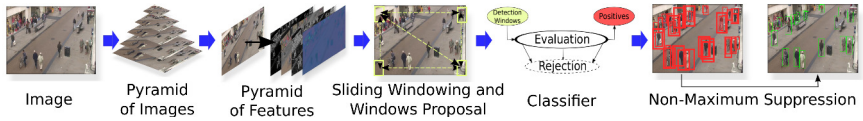


Figure 1.1: Standard execution pipeline of a multi-scale person detector.

computational cost of classification at each detection window [6, 19, 42, 82]. A few other researchers have proposed strategies to reduce the number of windows by selecting only those that are more likely to contain the target and obtain in this way a significant performance speed-up [12, 33, 57, 73]. Although all these methods have proved to be effective in reducing the computational effort, none of them has considered the opportunity of exploiting scene information (i.e. the statistics of the persons in the scene) that is typically available and usefully exploitable in most real contexts. In fact, in most of the real cases cameras are installed in fixed positions and observe a part of a scene. So, it is possible to learn a prior of the scene.

1.1 Contributions

We present in this thesis different solutions to obtain user interest profiling. The contribution of this work is divided into two main themes. The first one is related to the person detection and its complexities, while the second one deals the gaze estimation problem in low resolution image, without the need of a tracker. The thesis is organized as follow.

In Chapter 2 we review the state of the art methods for person detection and gaze estimation.

In Chapter 3 we present an approach to automatically improving the efficiency of the soft cascade-based person detectors, which addresses in the two fundamental bottlenecks in cascade detectors: the number of weak classifiers that need to be evaluated in each cascade, and the total number of detection windows that must be evaluated.

Following this latter line of research, in Chapter 4 we discuss an extension with a scene-dependent windows proposal method that grounds on gaussian mixture modelling of locations and scales of the persons in a scene.

Chapter 5 deals with the particular problem of person gaze estimation

in images and videos. We propose a coarse gaze estimation which can be exploited for video surveillance, for the analysis of social behavior interaction and for attention profiling. The solution exploits frame-to-frame motion information and therefore does not need to track every person in the scene, or perform complex and computationally onerous global optimization requiring the knowledge of the entire person trajectory.

In chapter 6 we present a real working installation opened in February 2015 at the National Museum of Bargello, namely Mnemosyne. Based on passive observation of the tourists during the visit, Mnemosyne extracts a profile of interests for each visitor and provides contextual information, author biography, related artworks and sites, based on the user's preferences.

In appendix A we describe the MuseumVisitors dataset [10] for person and group behavior understanding on which tracking, detection and coarse gaze estimation can be evaluated. All frames are recorded at National Museum of Bargello in Florence, Italy. The dataset provides camera calibration, object locations, annotation of groups, identities and occluded parts. The dataset has been recorded across different times of the day thus generating challenging sequences in term of lighting conditions.

Finally, in appendix B we present a Web Annotation Tool for Surveillance Scenarios (WATSS), developed to annotate the MuseumVisitors dataset. WATSS allows multiple users for concurrent annotations, with the possibility to insert groups and people identities, gaze and body occlusion.

Chapter 2

Literature review

This chapter gives a brief survey of related work on person detection and coarse gaze estimation. In the first part of the chapter, we review the state of the art of methods to speed-up the execution pipeline of a multi-scale person detection, while the second part deals with the problem of coarse gaze estimation on surveillance like scenarios, where head and body imagery are usually low resolution.

2.1 Multi-scale person detection

Recently many techniques have been proposed that improve the detection process both in terms of accuracy and efficiency. These methods can be roughly grouped based on the domain on which they act: the multi-scale feature representation, the method used for proposing detection windows or exploiting scene geometry, and the classifier used.

Notable solutions that expedite the computation of features in the detection window were proposed in [14, 34, 35, 37, 98]. In particular, in [35] the authors proposed the Haar-based Integral Channel Feature (ICF) for integral images that combines different types of features in an efficient way. This results in a reduction of computational effort with no loss of accuracy in the detection process. An approximation of ICF was proposed in [34] where features are extracted only at the middle-level of each octave and the nearby scales are obtained by interpolation. This approximation was exploited by Benenson et al. [14] in combination with three classifiers trained

on the three octaves of the pyramid. This moves the complexity of feature extraction from test to training time. A generalization of ICF was proposed by [98] that uses different filters bank instead of rectangular filters. All these methods however do not achieve a significant speed-up in the whole detection process although they proved to maintain good discriminative power. In order to reduce the cost of feature extraction in the pyramid, Dollár et al. extended the ICF into the Aggregate Channel Feature (ACF) [37], where distinct channel features obtained from block of pixels are aggregated.

Other researchers proposed new classifier architectures that perform effective classification at reduced computational cost [19,42,82]. Bourdev et al. proposed the Soft Cascade [19] where detections are evaluated at each node by taking into account also the weak-classifiers responses at the previous nodes. This permits early rejection of false positives, so reducing classification time while improving detection rate. In [82], an entropy-based rejection criterion was introduced in the Soft Cascade in order to allow early stopping of the evaluation of negatives. Felzenswalb et al. used the part-based deformable models (DPM) in conjunction with a cascade architecture to improve the detection accuracy [42]. Although all these methods achieve a good trade-off between accuracy and speed-up, nevertheless none of them has real-time performance. Solutions that exploit neural networks were proposed in [6, 26, 83]. In particular, in [83] a deep model was used to learn a discriminative representation of a person considering both person attributes and scene attributes. In [26], a two-stage system was proposed. In the first stage, the ACF detector is used to filter out the negative windows. Positive windows are then evaluated in the second stage with a trained Deep Convolutional Neural Network to obtain a feature that is used for classification. In [6], the feature cascades of [14] were used in a Deep Neural Network to speed-up the classification. All these methods require GPU implementations.

Solutions that avoid the sliding window over the full pyramid and use sparse windows proposals were presented in [51, 64, 75, 85, 93, 100]. These solutions reduce the number of detection windows to be evaluated either exploiting appearance properties and segmentation or adding constraints on the positions and scales of the target. In particular, in [51], the authors evaluate only the salient regions in the image. They are identified according to three distinct criteria: the visual contrast (*uniqueness*), the degree of blur (*focusness*) and the likelihood of having full visibility of the target (*objectness*). In [75], the initial candidate windows are obtained based

on an objectness prior considering the distribution of edges in the image. Then, efficient cascades are learnt to estimate the importance of the selected windows. Selective search was used in [85] to generate the most likely object locations based on hierarchical grouping. At each iteration region-level similarity is used to merge sub-regions of the image considering both texture and size information. In [64], to estimate salient regions, the authors merged superpixels based on a connectivity graph using partial spanning trees. The bounding box associated to each tree represents the location of a candidate window. In [100], salient regions were generated considering the edges extracted in the image and the importance of each window candidate was defined by the number of contours inside its region. In [93], superpixels were hierarchically grouped by considering the homogeneity of subregions. Although these sparse windowing strategies allow to reduce the number of detection windows to evaluate, none of them executes in real-time.

A priori information about the geometry of the scene was also used. In particular, in [12], the authors used stereo information to estimate stixels (all objects in the scene are described as vertical flat sticks on the ground) and determine targets height. In [8], the authors reduced the number of stages to be evaluated in the Soft Cascade and used spatial and scale statistics of persons in the scene. However, these solutions assume strong geometrical constraints, such as the vanishing line within the camera field of view [12] or a fixed grid superimposed on the image view [8].

Scene-specific person detectors have also been proposed [48, 50, 65, 89]. In [89], the authors introduced a transfer learning framework to adapt a generic person detector to a specific scene. To extract the training sets they considered both motion and scale information of the targets along with a path model. In [50] a verification strategy based on short-term tracking have been used to generate an accurate training set from the scene and train a classifier. A similar solution have been proposed in [65] that exploits target tracklets. Synthetic projections of persons on the image plane according to the geometry of the scene have been exploited in [48] to train a location-specific person detector. These solutions do not reduce the computational complexity of the detection process but focus on the automatic extraction of positive and negative samples to train a scene-specific classifier.

2.2 Coarse Gaze Estimation

Gaze and attention analysis are central topics in computer vision. In particular, gaze is usually inferred through head pose estimation which is in turn estimated by exploiting fast and accurate methods to detect stable face landmarks [55, 76, 94]. An even preciser gaze estimate can be computed by locating pupils inside eyeball regions [99]. However, all these methods require a fairly good resolution to obtain a reliable landmark estimation thus considering faces not smaller than 200 pixels. In visual surveillance scenarios, even if high resolution cameras are employed, it is often infeasible to obtain such resolution for all the faces of interest. Moreover, landmark and eye-detection based methods require frontal or profile faces to work, while persons are evenly imaged frontally or from their back.

For these reasons a different line of research tackled the relaxed problem of coarse gaze estimation [15, 17]. Instead of deriving a full 3D transformation for the head, coarse gaze estimation sets the goal of predicting the 2D orientation of the head with respect to the camera. For calibrated cameras such gaze can also be projected onto the scene ground plane [78].

Gaze prediction can be improved considering cues other than face imagery. Benfold *et al.* make the point that a gaze model is also context dependent, and propose an unsupervised model for learning scene-specific classifiers [16]. Another very relevant cue is obtained from the body orientation. Indeed the torso orientation poses a very strong constraint on the possible gaze angles. Moreover, if a person is in motion, the walking direction, which can be already used as weak predictor is also extremely relevant. Chen *et al.* learn body-head and velocity-head coupling factors [25]. Their approach is shown to improve with respect to [16]. However, both these approaches exploit a temporal model and therefore need a reliable tracker. Multi-target tracking is a very challenging task that can also be prone to failure in case of crowded environment. Moreover, being tracking the first block of a processing chain, its failure may lead to inconsistent results.

Chapter 3

Unsupervised scene adaptation for faster multi-scale pedestrian detection

In this chapter we describe an approach to automatically improving the efficiency of soft cascade-based person detectors. Our technique addresses the two fundamental bottlenecks in cascade detectors: the number of weak classifiers that need to be evaluated in each cascade, and the total number of detection windows that must be evaluated. By simply observing a soft cascade operating on a scene, we learn scale specific linear approximations of cascade traces that allows us to eliminate a large fraction of the classifier evaluation. Independently, we learn a coarse geometric model of the scene that allows to reduce the number of candidate windows run through the cascade. Both of our approaches are unsupervised and require no additional labeled person images for learning.

3.1 Introduction

Person detection provides the basic measurement model for tracking and person re-identification and is therefore a fundamental component of most modern surveillance systems. However, due to its computationally onerous

nature it is also the bottleneck in many systems. The general problem of detection has emerged as one of the major themes of modern computer vision research. Person detection in particular is an highly active topic of research. It has received a lot of attention in recent years, but remains an extremely difficult problem.

Person detection in unconstrained scenes is computationally expensive for several reasons. First of all, without knowledge of the geometry of the scene, every location and scale must be scanned for potential detections. Second, in the *soft cascade* detection architecture, currently the state-of-the-art for efficient person detection, a cascade of *weak classifiers* must be evaluated at each of these locations and scales to obtain a detection score. These two factors conspire to render unconstrained detection computationally onerous.

Improvements in the computational cost of person detection often address only one of these factors and rely on supervision such as manual calibration of the camera. We believe that it is crucial in practice that both factors be addressed with only weak or no supervision. In this work we propose two approaches to scene adaptation for soft cascade pedestrian detectors that need only to observe an already trained detection on the scene of interest. Our first adaptation strategy performs linear cascade approximation to avoid evaluating all stages of the soft cascade, while our second strategy minimizes the number of candidate windows evaluated using a statistical model of scales and position of likely detections in the scene.

In the next section we discuss the state-of-the-art in person detection. In section 3.3 we describe the soft cascade detection architecture which represents the current state-of-the-art. We describe our approach to learning how to detect faster in section 3.4, and in section 3.5 we report on a number of experiments we performed to evaluate our approach. We conclude in section 3.6 with a discussion of our contribution.

3.2 Related Work

Most state-of-the art methods follow the pipeline depicted in figure 3.1. Recently many techniques have been proposed that improve the detection process both in terms of accuracy and efficiency. These methods can be roughly grouped based on the domain on which they act: the multi-scale feature representation, the method used for proposing detection windows or exploiting scene geometry, and the classifier used.

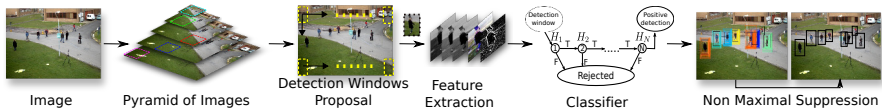


Figure 3.1: Standard execution pipeline of a multi-scale pedestrian detector. Given an image I, a pyramid is computed from it by progressively sampling by a fixed factor to obtain the set of levels. For each level are selected the detection windows and then from each of these are extracted the feature that will be considered by the classifier. Finally, for all detection windows not rejected, a non maximum suppression process is performed to obtain the final positive detection windows.

An approach to computational saving in the feature domain was proposed in [35]. The Integral Channel Feature for integral images uses a combination of different heterogeneous information channels to speedup the detection process while maintaining high accuracy. While a feature pyramid is mandatory for multi-scale detection, the authors of [34] proposed an approximation that avoid the direct computation of all levels of the feature pyramid by extracting them only for the median layer of each octave and approximating the remaining scales. This approximation takes the form of an exponential function that depends both on the type of the feature and on the position of the level in the octave. However, this preserves detector robustness only for an octave. In [14] the authors exploit a trained classifier for each octave and the approximation in [34] to avoid the computation of the features for each level in the octave.

Several methods have been proposed to speed up the computation by reducing the number of detection windows evaluated. In [21,47] the authors propose to first compute a sparse set of detector responses and then sample more densely around promising locations. In [33] the Crosstalk Cascade was proposed to simultaneously evaluate multiple candidates at a time exploiting two type of cascade: *excitatory* cascades that encourage a detection window with a neighborhood of possible positive responses and *inhibitory* cascades that reject detection windows with low partial scores in the neighborhood.

The geometry of the scene is also extensively exploited in the to speedup the computation and improve detection accuracy. For example, the method proposed in [31] exploits a calibration of the scene to improve and speedup a person detector by spatially filtering detection windows based on the ex-

pected height of a person. In [49] the authors propose a probabilistic inference model to merge pre-trained detector responses with scene geometry knowledge. However, this method requires that the vanishing lines be always visible in the image in order to estimate a coarse camera viewpoint from objects in the scene. The Stixels model used in [14] exploits a stereo vision system to extract depth information of the scene and then reduce the set of candidate detection windows.

In the classifier domain the Hard Cascade [86] improves both the accuracy and efficiency of the classic AdaBoost algorithm [87] by specializing the first stages in order to reject the majority of the negatives detection windows. The authors of [19] proposed the Soft Cascade architecture in which the evaluation of each detection window depends on the sum of all stages partial scores up to the current stage. The rejection threshold at each stage is learned considering the ROC surface, thus taking into account conjointly the speedup, the detection rate and the false positive rate.

We propose a framework to speed up the detection process by acting both in the classifier domain and in the scene geometry domain. The result is a significant reduction in the total number of stages evaluation required in the soft cascade detection process. To do this we exploit the *regions of support*, which refers to the suppressed positive detections that occur around a local maxima, to improve detector efficiency in:

- the classifier domain through linear approximation of soft cascades in order to estimate a final detection score without calculating all stages;
- the pyramid domain by locally modeling the scene-dependent statistics of detection windows and their scale distribution in order to focus effort on the evaluation of detection windows that are more likely to be a local maxima in the image.

Our approach does not require any *a priori* information about the scene and all learning is done by mining statistics about the soft cascade detector operating on a scene.

3.3 Pedestrian detection with soft cascade

In figure 3.1 we show the standard pipeline for person detection. Since the process of capturing an image from a scene can introduce changes in the scale of a pedestrian, a multi-scale detector is required. This is usually performed

by constructing a pyramid of images, which is a set of images obtained by progressively upsampling and downsampling the original image (referred to as the *levels* of image pyramid). Then each level is processed to extract the features. In particular, candidate regions are usually obtained using a sliding window at a fixed step size over all image levels. A classifier is then applied to each window for each level to assign a score. Finally, non maximum suppression is performed on positive candidates to obtain the final detection windows.

3.3.1 Multi-scale detection complexity

Without any optimization strategies, the evaluation of the whole pyramid of images in terms of total number of detection windows can be very expensive. Let L be the total number of levels of the pyramid, with m levels per octave, extracted for an image of $n \times n$ pixels, then the total number of windows that must be evaluated is:

$$\begin{aligned} \sum_{l=0}^{L-1} \mathcal{O}(n^2)\ 2^{\frac{-2l}{m}} &\approx n^2 \sum_{l=0}^{L-1} (4^{-\frac{1}{m}})^l \\ &= n^2 \left(\frac{1 - 4^{-\frac{L}{m}}}{1 - 4^{-\frac{1}{m}}} \right) \end{aligned} \tag{3.1}$$

Note that eq. (3.1) converges to $n^2/(1 - 4^{-\frac{1}{m}})$ for $L \to \infty$. Thus, for an image of 640×480 pixels with a pyramid of 3 octaves of 8 levels each, a total of $285,944$ detection windows must be evaluated.

3.3.2 The soft cascade classifier

An evolution of the cascade classifier used in [86] is the *Soft Cascade* proposed in [19]. To train a Soft Cascade, a set of rejection thresholds is learned in order to perform early stopping during the evaluation of negative detection windows. Given the feature vector $x \in \mathbb{R}^D$ of a sample detection window, and let $H : \mathbb{R}^D \to \mathbb{R}$ be a classifier composed of T stages, where each stage is a function $h_i : \mathbb{R}^D \to \mathbb{R}$. The partial score up to stage t is computed as:

$$H_t(x) = \sum_{i=1}^{t} h_i(x). \tag{3.2}$$

Let $\{\tau_t\}$ be the set of rejection thresholds, x is classified as *positive* with score $H_T(x)$ if $H_t(x) \geq \tau_t \quad \forall t \in [1, T]$. In this way the evaluation of each sample depends also on the scores obtained in the previous stages. Thus, considering the number of detection windows estimated in (3.1), it follows that using a soft cascade with 1024 stages requires the evaluation of approximately 10^9 stages for a single second of a video at 25 fps. This enormous number of cascade stages evaluated renders real-time pedestrian detection extremely challenging.

Figure 3.2: The Region of Support (ROS) around strong detections (black detection window) on a frame extracted from *Oxford*. The windows inside the same ROS have the same color and at the top-left of each strong detection window we report the cardinality of each ROS.

3.4 Unsupervised scene adaptation of soft cascade detectors

To avoid the computation of a very high number of stages as described in section 3.3 we propose a strategy to reduce the entire process by acting on both the classification and the detection windows proposal on the pyramid. The first contribution regards the total number of weak classifiers that must be evaluated to obtain a score for each positive detection window. In particular, we propose a solution to approximate the final score of a detection window without considering all the stages of a soft cascade. The second contribution provides an alternative strategy to the classic detection windows proposal that is able to avoid the sampling on the scene of those detection

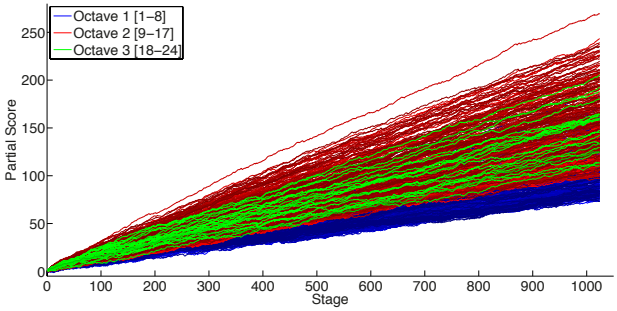


Figure 3.3: Positive traces extracted from soft cascade of 1024 stages on Oxford. The pyramid contains 24 levels (3 octaves of 8 levels each). Traces are colored based the level to which they correspond.

windows with a low probability of being a local maximum, in particular by filtering out those windows with a scale not consistent with the geometry of the scene. Both strategies are unsupervised and require only some frames extracted from the observed scene as a training set.

3.4.1 Leveraging region of support information

As reported in [33], the responses of the classifier on near positions and scales of the pyramid are related. A region of support (ROS) represents the extension of the sub-regions of an image in which all the detection windows (with different scale) are classified as positives. In general a ROS is composed of many intersecting detection windows, each with a different score. The window with the highest score is called local maximum (*strong*) because it is the only one that will survive the non maximum suppression procedure. Figure 3.2 shows some *strongs* with their respective ROS extracted from a soft cascade on a frame from the Oxford dataset [16]. The ROS shown can be very indicative of both the detector precision and the scene geometry, as well as the targets location inside the scene. In fact, the cardinality of each ROS can be used as a estimate of true positive for a detection window since the objects with a low rank in the frame are often false positive, e.g. the garbage and the mannequins. The location and scale of *strongs* can be considered to learn a model able to describe the geometry and perspective of the scene. All this information are very discriminative and can be extracted at no additional cost during the non maximum suppression process.

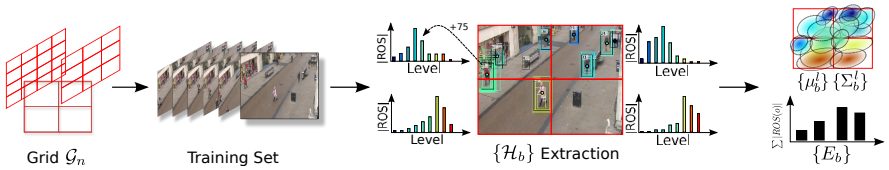


Figure 3.4: Pipeline for training the candidate window proposal model. After selecting the grid resolution, for each frame of the training set we extract the histogram of levels $\mathcal{H}_b$ and $\{\mu_b, \Sigma_b\}$ considering the ROI information of strong detections. Finally, for each block we estimate the energy parameter E_b to accentuate the research in sub-regions of the scene.

3.4.2 Linear cascade approximation

In figure 3.3 we plot the traces, that is the outputs of each weak classifier in the evaluation of a cascade, of many positive detections from the Oxford sequence. The trace plots are color coded according to which level in the image pyramid in which they were recoded. Note how all traces are basically linear. They are subject to local perturbations of limited energy, but the traces from the same level remain close to each other for their entire evolution. Considering this trend, we group the traces based on the level they come from and estimate a linear function that approximates the trend of curves from each level.

In particular, we define a linear score estimation function $\tilde{H}_{t\to T}(x) \in \mathbb{R}$ that requires the evaluation of only a fixed number $t < T$ of cascade stages and such that:

$$\tilde{H}_{t\to T}(x) \approx H_T(x), \tag{3.3}$$

where $H_T(x)$ represents the true cascade output obtained by evaluating all stages on input x. Given the trend observed in figure 3.3, we use linear regression and estimate the slope and intercept parameters for each trace. Formally this is obtained by solving the following minimization problem:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\| \mathbf{S}^{\mathbf{T}}\mathbf{w} - \mathbf{h}_{t\to T}(x) \right\| \tag{3.4}$$

where $\mathbf{w} \in \mathbb{R}^2$, $\mathbf{w} = [w_0 \quad w_1]$ with w_0 the intercept and w_1 the slope and with:

$$\mathbf{S} = \begin{bmatrix} 1 & \cdots & 1 & \cdots & 1 \\ t & t+\Delta & t+2\Delta & \cdots & T \end{bmatrix} \tag{3.5}$$

$$\mathbf{h}^{\mathbf{T}}_{t\to T}(x) = \begin{bmatrix} H_t(x) & H_{t+\Delta}(x) & \cdots & H_T(x) \end{bmatrix} \tag{3.6}$$

where Δ is the sampling step for the stages used in the regression. Under the maximum rank hypothesis of $\mathbf{S}$ the problem in Eq. (3.4) admits a unique solution $\hat{\mathbf{w}} = (\mathbf{S}\mathbf{S}^{\mathbf{T}})^{-1}\mathbf{S}\ \mathbf{h}_{t\to T}(x)$.

We compute the best parameters $\{\hat{\mathbf{w}}^l_i\}$ through eq (3.4) for each trace in each level l of the pyramid and then we estimate the final parameter $\overline{\mathbf{w}}^l$ by averaging on that. This is done for each level of the pyramid. The final score approximation is:

$$\tilde{H}_{t\to T}(x) = \overline{\mathbf{w}}^l \cdot [0 \quad T-t\,] + H_t(x) + \bar{\epsilon}^l \tag{3.7}$$

where l is the level of x, ϵ^l is an error obtained as $E[H_T(x) - (\overline{\mathbf{w}}^l \cdot [0 \quad T - t\,] + H_t(x))]$ for $x \in V$, and V is a validation set. Note that Eq. (3.7) does not consider w_0, since the approximation is constrained to pass through $H_t(x)$ in that $\tilde{H}_{t\to t}(x) \simeq H_t(x)$.

Eq. 3.7 is easy and fast to compute and can be used to obtain an approximation of the final score of a detection window. This approximation requires the evaluation of only the first t stages of the soft cascade. Note also that it is completely unsupervised in that we only require a sample of cascade evaluations from an already trained soft cascade detector and do not require additional labeled training data to fit the model parameters.

3.4.3 A generative model for candidate window proposal

The naive soft cascade approach to detection achieves scale invariance by exhaustively scanning all locations and scales in an image. In practice, especially in typical surveillance scenarios using fixed cameras, not all scale/location combinations are feasible due to the geometry of the scene. Our second strategy is to learn a generative model for candidate window proposal in order to reduce the number of candidate windows extracted from the pyramid. We do this without relying on calibration or any additional information. As shown in figure 3.2, the presence and scale of targets is highly dependent on the geometry of the scene. Since the geometric information of the scene is directly related to the level of the pyramid, we argue that the complete evaluation of all possible levels of the pyramid in all sub-regions of the image is wasteful. Instead, we will exploit the ROS for observed strong detections (i.e.

those that survive non maximum suppression) in order to propose candidate windows for each scale and position combination in the scene.

Learning the generative model The pipeline of the proposed model is shown in figure 3.4. To extract the statistics of subregions of the scene we divide each frame of the training set into $n \times n$ rectangular blocks. Inside each block b, the strong detections observed in the training set are used to compute a histogram $\mathcal{H}_b$ where each bin $\mathcal{H}_b^l$ represents a level of the pyramid. Specifically, the strong detections in the block b contribute with the cardinality of their ROS in the corresponding bin level. The cardinality of the ROS is the number of detections that are suppressed by the overlapping strong detection. This provides a robust local description of the frequent scales in a block.

To extract information about the representative locations in a block for a certain level we thus compute the average centroid position μ_b^l and its covariance Σ_b^l on the strong detections. This is useful to estimate the real locations in the scene where person detections occur with high probability. Finally, for each block we compute an energy factor E_b, such that:

$$E_b = \frac{\sum_{l=1}^{L} \mathcal{H}_b^l}{\sum_{\tilde{b} \in \mathcal{G}_n} \sum_{l=1}^{L} \mathcal{H}_{\tilde{b}}^l}, \tag{3.8}$$

where $\mathcal{G}_n$ indicate the set of blocks. This factor emphasizes the research for certain sub-regions by generating the detection windows proportionally. The final model is:

$$M_n = (\mathcal{G}_n, \{\tilde{\mathcal{H}}_b^l\}, \{\mu_b^l, \Sigma_b^l\}, \{E_b\}), \tag{3.9}$$

where $\tilde{\mathcal{H}}_b^l$ indicates $\mathcal{H}_b^l$ normalized over all levels in block b.

Candidate window proposal at detection time The number of detection windows of a pyramid to be evaluated is chosen proportionally to a parameter $\gamma \in [0, 1]$. This parameter is used as an estimate of the final speedup we want from the resulting detector. There is clearly a tradeoff between speed ($\gamma \to 0$) and accuracy ($\gamma \to 1$) of the detector. In particular, given a test frame I, the number of detection windows that we evaluate for each block b and level l in the pyramid $\mathcal{P}(I)$ is:

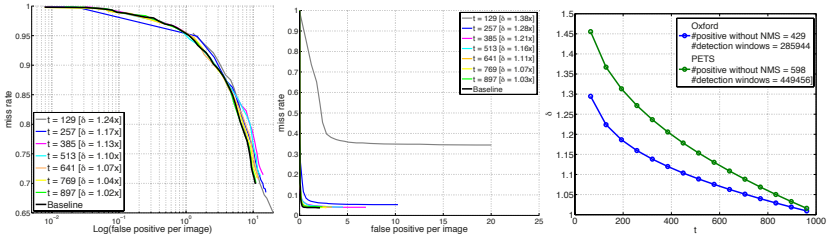
$$N = \gamma \; |\mathcal{P}(\mathcal{I})| \; E_b \; \tilde{\mathcal{H}}_b^l. \tag{3.10}$$

where $|\mathcal{P}(\mathcal{I})|$ corresponds to the total number of detection windows in pyramid $\mathcal{P}(\mathcal{I})$.

At detection time we sample detection windows using an iterative procedure. In the first iteration we randomly sample N detection windows from the normal distribution $\mathcal{N}(\mu_b^l, \Sigma_b^l)$. From this set of detection windows we remove duplicates and if necessary perform another iteration, expanding the covariance matrix by a fixed factor s along the principal directions of the covariance matrix Σ_b^l.

Note that this strategy for improving the efficiency of soft cascade detection is also completely unsupervised. We build our scale- and position-local generative models by analyzing the behavior of strong detections and their regions of support on a training set of detection outputs. At detection time we can control the number of candidate windows proposed and thus control the efficiency/accuracy tradeoff of the final detector.

3.5 Experimental results



(a) (b) (c)

Figure 3.5: ROC curves of baseline using the linear cascade approximation, for different values of t, in sequence Oxford (a) and PETS (b). In bracket we show the obtained saving. (c) Saving (delta) for different values of t, using the linear cascade approximation. The maximum reduction is under the 40% (1.5x).

In this section we report on the performance for the proposed approach using the linear cascade approximation, the candidate windows proposal model, and the combination of both. We use the soft cascade detector implemented in the OpenCV repository[1] as a baseline. We use two datasets in our experiments: Oxford [16] and PETS [4]. The Oxford dataset is a

[1] `https://github.com/Itseez/opencv`

challenging full HD video sequence due to high variation of pedestrian scale, occlusions and confusion with shopping window mannequins. For the PETS dataset we considered the *s2.l1-view1* sequence with an image resolution of 768×576 pixels. We extracted 180 frames from Oxford by sampling one over fifteen frames and 199 frames from PETS by sampling one over four frames. From these frames we a third for the training and the remaining for the test. All comparisons between different detectors are given using ROC curves in terms of miss rate versus false positive per image. The baseline is represented by the soft-cascade with 1024 stages using a classifier for each octave and a pyramid of images consisting of 3 octaves of 8 levels each.

The performance of our proposed approaches is measured as function of a savings factor δ that is computed as:

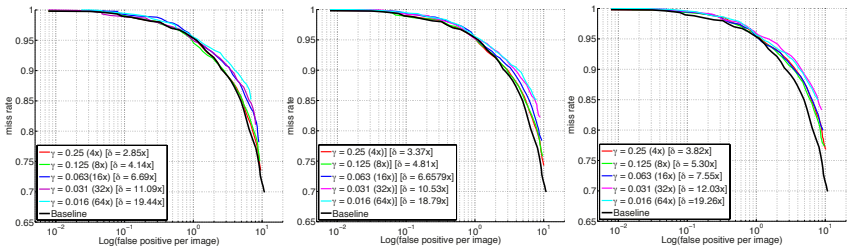
$$\delta = \frac{\sum_{\forall x \in \mathcal{P}} [H(x)]}{\sum_{\forall x \in \mathcal{X}} 1_{\{c=0\}} [H(x)] + 1_{\{c=1\}} [\tilde{H}_{t \to T}(x)]} \tag{3.11}$$

where the operator $[\cdot]$ returns the number of stages computed, c indicates if the linear cascade approximation is used ($c = 1$) and $\mathcal{X} = \mathcal{P}$ when all sliding detection windows are considered or $\mathcal{X} = \tilde{\mathcal{P}}$ when the set of detection windows is obtained from our generative model for candidate window proposal.

3.5.1 Experiments with linear cascade approximation

In this section we analyze the performance of linear cascade approximation for different t values. Observe in Figure 3.5(a) how on the Oxford sequence, the curves of the proposed approximation are close to the baseline, with a gradual reduction in loss when the number of stages evaluated increases. The total savings varies from 19% (1.24×) with 129 stages to 2% (1.02×) with 897 stages evaluated. For the PETS sequence, shown in figure 3.5(b), loss is drastically reduced for $t > 129$ stages. The maximum saving reached with this sequence is 28% (1.38×).

In figure 3.5(c), we show the savings evolution varying the number of stages evaluated for both sequences. Considering a small number of stages for each detection window, the computational savings is at most 23% (1.3×) in Oxford and 31% (1.45×) in PETS. The savings is modest because the computational cost is mostly dominated by the total number of *negative* windows evaluated, that decreases exponentially with increasing t (the number of stages considered for the linear cascade approximation). Linear cascade



(a) Grid 2×2 (b) Grid 4×4 (c) Grid 6×6

Figure 3.6: ROC curves using candidate window proposal on Oxford sequence for a range of γ and grid sizes 2×2 (a), 4×4 (b), 6×6 (c).

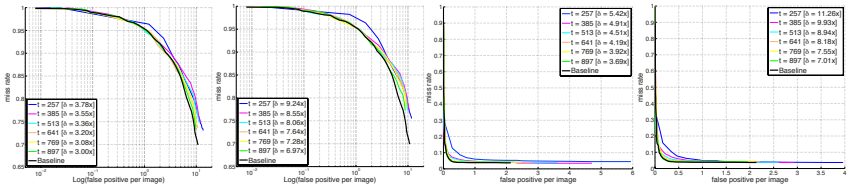
approximation helps, but to achieve significant computational cost reduction the total number of the candidate windows must be reduced.

3.5.2 Experiments with candidate windows proposal

We evaluated the performance of our candidate window proposal model on the Oxford sequence for different values of γ and grid dimensions. The results are shown in figure 3.6. Each plot shows results for different grid resolutions (2×2, 4×4 and 6×6) and varying the speedup parameter γ. In general, with all configurations we obtain a savings greater then 50% ($2\times$). For example, for a grid size of 2×2, the minimum and maximum saving values is 65% ($2.85\times$) and 95% ($19.44\times$), respectively. Considering the savings in computation, the loss in accuracy with respect to the baseline is very low at 10^{-1} *fppi* (under 0.5%). Increasing the grid resolution results in a small performance drops with respect to the baseline. The grid 2×2 is the best configuration in terms of loss and savings. This is due to the fact that, despite the large blocks in the 2×2 grid configuration, covariance expansion will ensure that the Gaussian will still eventually cover the whole block.

3.5.3 Experiments with both strategies

In this section we evaluate the combination of both proposed strategies on the Oxford and PETS sequence (see figure 3.7). Results are shown for different values of γ and t with a grid resolution of 2×2. On Oxford, with $\gamma = 0.25$



(a) $\gamma = 0.25$ (4×) (b) $\gamma = 0.0625$ (16×) (c) $\gamma = 0.25$ (4×) (d) $\gamma = 0.0625$ (16×)

Figure 3.7: ROC curves for both strategies, with a grid size of 2×2 and $\gamma \in \{0.25, 0.0625\}$, on both the Oxford (a-b) and PETS(c-d) sequences.

(4×) the maximum savings obtained respect to the baseline is 74% (3.78×), 9% more than the candidate window proposal alone, with no loss. For PETS, with $\gamma = 0.25$ (4×) we obtain a reduction of 81% (5.42×) with respect to the baseline, while with $\gamma = 0.0625$ (16×) we reach the 91% (11.26×) of saving. With both values of γ and $t \geq 513$ the obtained curves are the best in terms of accuracy with respect to the baseline, with a loss under 5%. The combination of the proposed strategies result in higher savings compared to the candidate windows proposal strategy while sacrificing little in terms of accuracy.

3.5.4 Comparison with the state-of-the-art

A direct comparison with state-of-the-art techniques is difficult due to the unavailability of source code for detectors, differences in protocols for evaluation, and ambiguities in how speedup is measured. Nonetheless, we make here qualitative observations about the performance of our approach with respect to the Crosstalk Cascade [33], which is similar in spirit to our candidate window proposal strategy. They evaluate performance on an images dataset, and thus we cannot directly compare on the same data since we require video on which to learn our model parameters. Losses of 0.1%, 0.5% and $2-4$% for savings of 4×, 8×, and $16-32$×, respectively, are reported in [33]. Adjusting the γ and t parameters of our candidate proposal model to match these savings as closely as possible, we obtain losses of -0.4%, 0.1%, and 0.5% for savings of 3.78×, 8.06×, and 25×, respectively.

3.6 Conclusion

In this work we proposed two strategies to reduce the computational complexity of a multi-scale pedestrian detector. Both strategies are unsupervised, based only on region of support information measured on a training set of unlabeled images. Our experiments demonstrate that both techniques are effective at increasing the efficiency of detection while sacrificing little in terms of accuracy. Linear cascade approximation yields modest improvement in efficiency due to the fact that the evaluation of negative windows dominates the total computation time. Candidate window proposal instead yields significant gains since it reduces the total number of candidate detection windows considered.

Chapter 4

Scene-dependent Windows Proposals for Efficient Person Detection

In this chapter we extend the strategy proposed in Chapter 3 with a new method for scene-dependent windows proposal that provides a substantial speed-up of person detection while showing high classification accuracy. This method learns a Gaussian Mixture Model of locations and scales of the persons in the scene under observation. The model is learnt from a set of detections extracted from a small number of frames. The mixture distribution is learnt in an unsupervised way so that each component of the mixture represents the expectation of finding a target in a region of the image at a specific scale. At runtime, the windows that most likely contain a person are sampled from the components and evaluated by the classifier. Experimental results show replacing sliding windows with our scene-dependent windows proposal in state of the art person detectors allows us to drastically reduce the computational complexity while granting equal or superior performance in terms of accuracy.

4.1 Scene-dependent windows proposal

While developing on some ideas of our previous proposed method [8], in this work, we present an innovative solution for scece-specific windows proposals that grounds on Gaussian Mixture Modelling (GMM) of locations and scales of the persons in a scene. The GMM is learnt in an unsupervised way from a set of detections extracted from a small number of frames of the scene. Each component of the mixture is a probability distribution that represents the expectation to find the target in a region of the image at a specific scale. In the GMM fitting procedure, any detection is weighed according to its importance in its Region of Support (ROS) [33]. At runtime, the windows that most likely contain a person are sampled from the components of the mixture, and evaluated by the classifier.

4.1.1 Scene model representation

The scene is modelled as:

$$p(y_i|\boldsymbol{\theta}) = \sum_k \hat{\alpha}_k \, p(y_i|\hat{\boldsymbol{\theta}}_k), \tag{4.1}$$

$$p(y_i|\hat{\boldsymbol{\theta}}_k) \sim \mathcal{N}(y_i|\, \boldsymbol{\mu}_k, \boldsymbol{C}_k)$$

where y_i is a generic target detection in the scene and $\boldsymbol{\theta} \equiv \{\hat{\boldsymbol{\theta}}_1, ..., \hat{\boldsymbol{\theta}}_K, \hat{\alpha}_1, ..., \hat{\alpha}_K\}$ includes the mixture components $\hat{\boldsymbol{\theta}}_k$ of the GMM and their mixing probabilities $\hat{\alpha}_k$ (both to be estimated). Each component $\hat{\boldsymbol{\theta}}_k$ of the GMM has mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{C}_k$, i.e.:

$$\hat{\boldsymbol{\theta}}_k = \{\boldsymbol{\mu}_k, \boldsymbol{C}_k\}, \quad \boldsymbol{\mu}_k \in \mathbb{R}^2, \, \boldsymbol{C}_k \in \mathbb{R}^{2\times 2}$$
$$\hat{\alpha}_k \in \mathbb{R}, \quad \sum_k \hat{\alpha}_k = 1$$

and represents with mixing probability $\hat{\alpha}_k$ a region of the scene with probability of having targets at a certain scale. Parameter $\boldsymbol{\mu}_k$ and $\boldsymbol{C}_k$, represent respectively the position of the centroid and the extension of a region of the image.

The GMM is learnt in an unsupervised way from the set of detections extracted with a person detector before Non Maximum Suppression, in a small number of frames of the scene. Fitting the GMM to the detections is

performed using a modified version of the Component-Wise EM (CEM) algorithm [43]. The conditional distribution $p(y_i|\,\hat{\boldsymbol{\theta}}_k)$ represents the likelihood that a detection y_i is generated by the GMM component $\hat{\boldsymbol{\theta}}_k$ and is defined as:

$$p(y_i|\,\hat{\boldsymbol{\theta}}_k) = \frac{s_{i,k}}{S}\,\mathcal{N}(y_i|\,\boldsymbol{\mu}_k, \boldsymbol{C}_k) \tag{4.2}$$

where $S = \sum_k s_{i,k}\,\mathcal{N}(y_i|\,\boldsymbol{\mu}_k, \boldsymbol{C}_k)$ is a normalization factor and $s_{i,k}$ is a multiplicative factor that accounts for the relative importance of y_i in relationship to the other detections in its ROS and the probability that the scale of the GMM component $\hat{\boldsymbol{\theta}}_k$ well fits with the scale of y_i, i.e.:

$$s_{i,k} = p\big(y_i|w(y_i)\big)p\big(y_i|\boldsymbol{\Sigma}(l(\hat{\boldsymbol{\theta}}_k))\big)$$

where:

$$p\big(y_i|w(y_i)\big) = \frac{w(y_i)}{\sum_j w(y_j)} \tag{4.3}$$

being $w(y_i)$ the relative weight of y_i, appropriately defined as in section 4.1.2.

$$p\Big(y_i\,|\boldsymbol{\Sigma}\big(l(\hat{\boldsymbol{\theta}}_k)\big)\Big) = \frac{1}{Z}\exp\{-\frac{\big(l(\hat{\boldsymbol{\theta}}_k) - l(y_i)\big)^2}{2\boldsymbol{\Sigma}\big(l(\hat{\boldsymbol{\theta}}_k)\big)}\} \tag{4.4}$$

being Z a normalization factor over the variances of the scales, $l(\cdot)$ a function that returns the scale and

$$\boldsymbol{\Sigma}(l(\hat{\boldsymbol{\theta}}_k)) = \frac{1}{M}\sum\big(l(\hat{\boldsymbol{\theta}}_k) - l(y_i)\big)^2$$

with M equal to the number of detections that fall in the ROS, such that their maximum has scale equal to the scale of the GMM component $\hat{\boldsymbol{\theta}}_k$.

GMM fitting is shown in pseudocode in Algortihm 1. Given the initial condition $\boldsymbol{\theta}$, appropriately defined as in section 4.1.3, at each iteration a new configuration $\boldsymbol{\theta}$ is estimated. The objective function $\mathcal{L}(\boldsymbol{\theta})$ is then applied:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) = \sum_{k=1}^{K}\frac{|\hat{\boldsymbol{\theta}}_k|}{2}\Bigg(log\Big(\frac{|\{y_i\}|\alpha_k}{12}\Big) + \frac{K}{12}log\Big(\frac{|\{y_i\}|}{12}\Big) \\ + \frac{K(|\hat{\boldsymbol{\theta}}_k|+1)}{2}\Bigg) - log\,p(\{y_i\}|\boldsymbol{\theta})\end{aligned} \tag{4.5}$$

where $(|\hat{\boldsymbol{\theta}}_k|/2)log(|\{y_i\}|\alpha_k)$ represent the optimal sample size for estimating θ_k and $-log\, p(\{y_i\}|\boldsymbol{\theta})$ represent the code-length of the data, according to the Minimum Description Length principle [46]. If $\mathcal{L}(\boldsymbol{\theta})$ is lower than the best configuration $\mathcal{L}_{best}$ obtained so far, the current components will replace the best configuration. A low value of $\mathcal{L}(\boldsymbol{\theta})$ means that the learned model well fit the detections windows $\{y_i\}$.

The least significant components of the GMM are discarded according to the *representativeness* factor $\hat{\boldsymbol{\beta}}_k$ that accounts for both *coverage* (i.e. how much a component is covered by another - components with a greater extension are promoted), and *overlapping* (i.e. how much a component overlaps with another - overlapped components are penalized), and defined as:

$$\hat{\boldsymbol{\beta}}_k = \frac{1}{2}\left(1 + \frac{Area(\boldsymbol{C}_k)}{\max_j Area(\boldsymbol{C}_j)} - \frac{Area(\bigcup_j \boldsymbol{C}_j - \boldsymbol{C}_k)}{Area(\bigcup_j \boldsymbol{C}_j)}\right). \tag{4.6}$$

(a) *Scales* (b) *TownCentre* (c) *PETS S2.L1-view1* (d) *CAVIAR*

Figure 4.1: GMM fitting on sample frames obtained with the VeryFast person detector [14]. Only the most relevant components of the GMM are shown.

This process is repeated until the number of components of the GMM is equal to a fixed number K_{min}. Figure 4.1 shows examples of GMM fitting for selected frames of sample sequences.

4.1.2 Detection weighting

Regions of Support are defined as the set of detections y_i that have intersection higher or equal to a predefined threshold τ_{voc}. Typically τ_{voc} is set to 0.65 [38]. In our previous work [8], we used a single weight for each ROS that

Algorithm 1 GMM Fitting

Input: Detection windows $\{y_i\}, \boldsymbol{\theta}_{\text{init}}, K_{min}$
Output: GMM with best configuration $\boldsymbol{\theta}$

```
L_best ← +∞
θ_best ← θ
while Number of components θ̂_k > K_min do
    [θ, L(θ)] ← CEM({y_i}, θ) with Eq. (4.2)
    if L < L_best then
        L_best ← L
        θ_best ← θ
    end if
    k* ← arg min_k { α̂_k · β̂_k }
    Remove the component θ̂_{k*} from θ
end while
θ ← θ_best
```

was equal to the number of the detections in the ROS (Hard ROS weighting). Such set of weights, learnt from the detections in a training set, was used to generate the window proposals in the same scene at runtime. As a modification of this approach we have assigned a weight for each detection y_i that accounts for its relative importance in the ROS in relationship to the other detections (Soft ROS weighting), i.e.:

$$w(y_i) = \begin{cases} \pi(y_i)\,\rho(y_i)\,|ROS(y_i)| & \text{if } y_i \text{ is not a local maximum} \\ |ROS(y_i)| & \text{otherwise} \end{cases} \tag{4.7}$$

where:

- $\pi(y_i)$ represents the relative *positiveness* of the detection window y_i defined as $g_i - \tau_T$ where g_i is the classifier score and τ_T is the rejection threshold of the cascade classifier at the scale of y_i, scaled with respect to the maximum in the ROS.

- $\rho(y_i)$ represents the *relevance* of y_i defined as:

$$\rho(y_i) = \left[1 + \exp\{-\frac{6}{(1-\phi)}\left(\frac{[\![y_i]\!] - \phi|ROS(y_i)|}{|ROS(y_i)|}\right)\}\right]^{-1}$$

  being $[\![y_i]\!]$ the rank in the ROS given by the classifier, and $\phi \in [0, 1)$

Dataset	Detections			Filtered Detections		
	Avg. n° detections/frame	Avg. Min-Max Scales	Avg. n° scales/frame	Avg. n° detections/frame	Avg. Min-Max Scales	Avg. n° scales/frame
TownCentre	341	6-19	13	31	6-18	12
PETS S2.L1-view1	443	3-19	16	50	3-19	15
CAVIAR	252	8-16	9	7	9-16	8

Table 4.1: Detections statistics with and without pre-filtering, using the VeryFast detector on test datasets.

is used to center the sigmoid, so that:

$$\rho(y_i) = \begin{cases} > 0.5 & \text{if } [\![y_i]\!] > \phi \cdot |ROS(y_i)| \\ 0.5 & \text{if } [\![y_i]\!] = \phi \cdot |ROS(y_i)| \\ < 0.5 & \text{otherwise} \end{cases}$$

Detections with $[\![y_i]\!] < \phi \cdot |ROS(y_i)|$ have lower weight so to inhibit the contributions of non informative detection windows ($\phi = 0.5$ is used).

4.1.3 GMM initialization

Deriving a GMM model from the full set of detections is generally unfeasible. To this end, for each image of the training sequence we apply the Otsu transformation [70] after the background subtraction to obtain the binary foreground mask and extract the foreground connected regions. Hence, only the detections that have overlap higher than 50% with such regions are retained. Table 4.1 reports the effects of pre-filtering on three test datasets. While the average number of detections per frame is drastically reduced, we can observe that the range of scales is preserved.

Given the set of filtered detections, for each scale of the pyramid, we iteratively execute the K-means algorithm decreasing the number of components at each iteration until the second moments of each cluster form a semi-definite positive matrix. This process is repeated separately for each scale and all the components are hence combined together to initialize the GMM model. The maximum number of clusters for each scale is chosen proportionally to the number of local maxima of the scale.

4.1.4 Mixture-based windows proposal

At runtime, the set of detection windows generated by the GMM is evaluated according to Algorithm 2. At each iteration, n detection window centroids

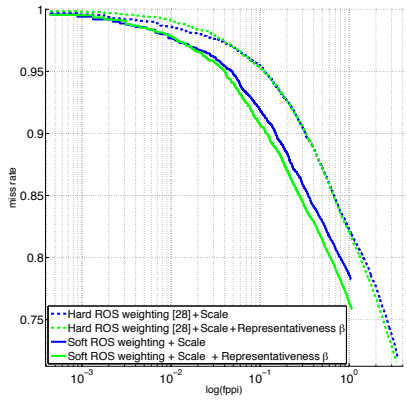


Figure 4.2: ROC curves of detection obtained with the VeryFast detector and our windows proposal (TownCentre dataset, $\gamma = 1/8$): effects of scale, representativeness β, Hard and Soft ROS.

Algorithm 2 Mixture-based Windows Proposal

Input: $\boldsymbol{\theta}, \gamma, N$
Output: set of sparse detection windows Y
  $Y \leftarrow \emptyset$
  for each $\hat{\boldsymbol{\theta}}_k$ **do**
    $n \leftarrow \gamma N * \frac{\hat{\boldsymbol{\alpha}}_k \, Area(\boldsymbol{C}_k)}{\sum_j \hat{\boldsymbol{\alpha}}_j \, Area(\boldsymbol{C}_j)}$
    $\{y_i\} \leftarrow$ Sample n detection windows using $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{C}_k)$
    $Y \leftarrow Y \cup \{y_i\}$
  end for

are sampled from the normal distribution $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{C}_k)$:

$$n = \gamma N * \frac{\hat{\boldsymbol{\alpha}}_k \, Area(\boldsymbol{C}_k)}{\sum_j \hat{\boldsymbol{\alpha}}_j \, Area(\boldsymbol{C}_j)}$$

where N is the maximum number of detection windows and $\gamma \in [0, 1]$ the fraction of windows that are passed to the classifier, as in [8]. This factor controls the speed-up of the detector and there is clearly a tradeoff with recall ($\gamma \longrightarrow 1$). Centroids are converted in detection windows with height corresponding to the scale of the Gaussian component.

	Scale	Scale + Representativeness
Hard ROS weighting [8]	0.95	0.95
Soft ROS weighting	0.92	0.91

Table 4.2: Miss-rate @10^{-1} FPPI of detection with sparse windows proposals (TownCentre dataset, $\gamma = 1/8$).

Figure 4.9 shows the performance improvement ascribed respectively to Soft ROS weighting (Eq. (4.7)), scale (Eq. (4.4)) and representativeness β (Eq. (4.6)) of detections in GMM fitting. Tests were conducted on the TownCentre dataset [16], with $\gamma = 1/8$ using the cascade classifier of the VeryFast detector [14] applied to the set of windows proposals. Soft ROS weighting with scale information are responsible for the highest improvement. This can be explained by the fact that Soft ROS weighting considers detections at any scale and these detections are re-weighed according to their scale. From Table 4.4 it can be observed that GMM fitting with weighting, scale and representativeness of detections for each ROS, defines a model that better generalizes the observed scene while granting a minimal coverage for each Gaussian component.

We will refer to our windows proposal method as *Mixture-based Windows Proposal* (MWP) in the following.

4.2 Experimental results

In this section, we report a set of experiments to assess the performance of MWP. Three public available datasets have been considered, namely: TownCentre [16], PETS S2.L1-view1 [3] and CAVIAR [1]. For the TownCentre

	TownCentre dataset				
	Baseline	$\gamma = 1/4$	$\gamma = 1/8$	$\gamma = 1/16$	$\gamma = 1/32$
HOG [30]	2.43 (1)	4.92 (1)	7.39 (1)	9.98 (1)	13.84 (1)
DPM [42]	0.23 (0.91)	- (0.86)	- (0.88)	- (0.89)	- (0.9)
ChnFtrs [35]	1.52 (0.96)	- (0.94)	- (0.95)	- (0.95)	- (0.95)
VeryFast [14]	11.61 (0.96)	13.95 (0.90)	22.15 (0.91)	31.9 (0.92)	40.65 (0.92)
ACF [37]	23.31 (0.97)	27.81 (0.96)	33.70 (0.96)	42.24 (0.97)	57.85 (0.97)
	PETS S2.L1-view1 dataset				
	Baseline	$\gamma = 1/4$	$\gamma = 1/8$	$\gamma = 1/16$	$\gamma = 1/32$
HOG [30]	3.34 (1)	5.67 (1)	8.25 (1)	9.31 (1)	11.22 (1)
DPM [42]	0.23 (0.56)	- (0.41)	- (0.46)	- (0.48)	- (0.49)
ChnFtrs [35]	1.51 (0.52)	- (0.42)	- (0.47)	- (0.48)	-(0.49)
VeryFast [14]	10.41 (0.36)	9.36 ($\sim$ 0.41)	15.01 ($\sim$ 0.42)	22.00 ($\sim$ 0.45)	30.47 ($\sim$ 0.47)
ACF [37]	22.34 (0.57)	24.93 (0.48)	32.15 (0.54)	39.48 (0.57)	46.19 (0.57)
	CAVIAR dataset				
	Baseline	$\gamma = 1/4$	$\gamma = 1/8$	$\gamma = 1/16$	$\gamma = 1/32$
HOG [30]	7.51 (1)	8.83 (1)	10.55 (1)	12.17 (1)	14.69 (1)
DPM [42]	0.23 (0.83)	- (0.8)	- (0.8)	- (0.81)	- (0.81)
ChnFtrs [35]	1.38 (0.86)	- (0.85)	- (0.85)	- (0.85)	- (0.86)
VeryFast [14]	11.95 (0.86)	13.03 (0.85)	21.31 (0.85)	31.18 (0.84)	40.01 (0.85)
ACF [37]	21.77 (0.86)	24.57 (0.84)	31.79 (0.85)	43.65 (0.85)	57.03 (0.86)

Table 4.3: Comparative performance of MWP against sliding windows for the tested person detectors. Number of frames per second of the full classification pipeline are reported with miss-rates@10^{-1} FPPI in parenthesis.

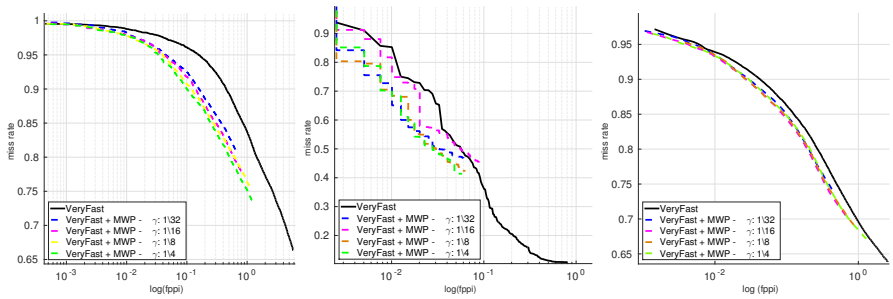
and the PETS S2.L1-view1 datasets we considered the first half of the images for training and the remaining for testing. For the CAVIAR dataset we considered the 26 clips of the *Corridor Viewpoint* for a total of 36293 frames and adopted the leave-one-out strategy (one clip is used as test while the other 25 are used for training) and evaluated the average accuracy. For each dataset, video frames were resized to 640×480 pixels.

We run five state of the art person detectors, namely: HOG [30], DPM [42], ChnFtrs [35], VeryFast [14] and ACF [37] on the training set to extract person detections. Those detections are used to learn our scene-depenedent windows proposal model, one for each dataset and person detectors.

At test time, for the HOG, VeryFast and ACF detectors we modify the original algorithm and replace the sliding windows (in the windows proposal stage) with our MWP. This was not possible for the ChnFtrs and DPM detectors and in this case, before the non-maximum suppression stage, we automatically filter out all those windows that do not overlap with the ones generated by our model.

All the experiments have been performed on an Intel Xeon@2.67 GHz (8-core) with 20 GB RAM.

Table 4.3 report the number of frames per second and miss rates at 10^{-1} FPPI obtained with sliding windows or with our MWP, for different values of γ. It is possible to observe that using MWP in a person detector allows obtaining similar miss rate performance, with respect to sliding windows, with much higher efficiency.

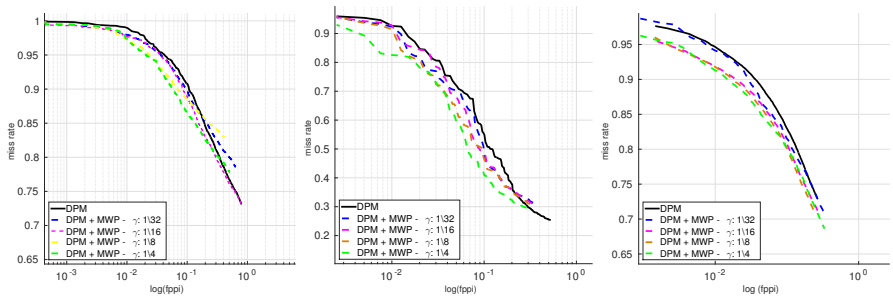


(a) TownCentre (b) PETS S2.L1-view1 (c) CAVIAR

Figure 4.3: Performance comparison on the TownCentre, PETS S2.L1-view1 and CAVIAR datasets using MWP with VeryFast detector

Fig 4.3 shows the ROC curves of the VeryFast person detector on the three tested datasets. On TownCentre, the use of MWP allows obtaining a significant reduction of the miss-rate, over 0.05 at 10^{-1} FPPI with respect to sliding the windows over the whole image. This is mainly due to the ability of our solution to select windows at the true scale of the target. For the PETS S2.L1-view1 dataset, MWP drastically limits the false positives generated by the Veryfast detector resulting in a better accuracy under 10^{-1} FPPI. No substantial improvement can be observed on the CAVIAR dataset due to the low resolution of the original video sequences. In terms of efficiency, MWP with the VeryFast detector allows improving the execution time from 12fps up to 40fps.

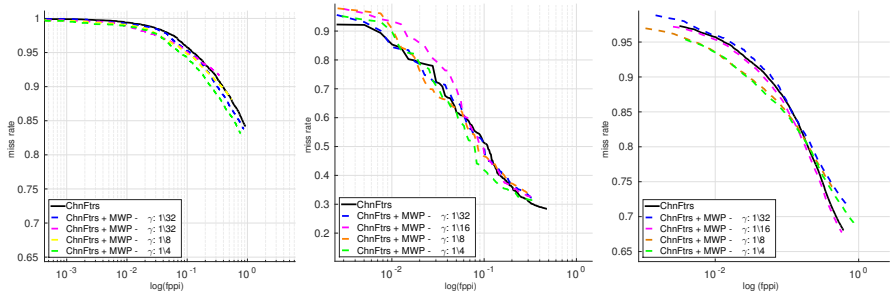
Fig 4.4 reports the performance obtained with the DPM detector. There is no significant difference in accuracy with MWP for $\gamma < 1/8$, while with $\gamma = 1/4$ our method achieves the best result at 10^{-1} FPPI on the TownCentre. The limited improvement of MWP in this case is motivated by the ability of DPM to manage the strong occlusions and the high scale variations that



(a) TownCentre (b) PETS S2.L1-view1 (c) CAVIAR

Figure 4.4: Performance comparison on the TownCentre, PETS S2.L1-view1 and CAVIAR datasets using MWP with DPM detector

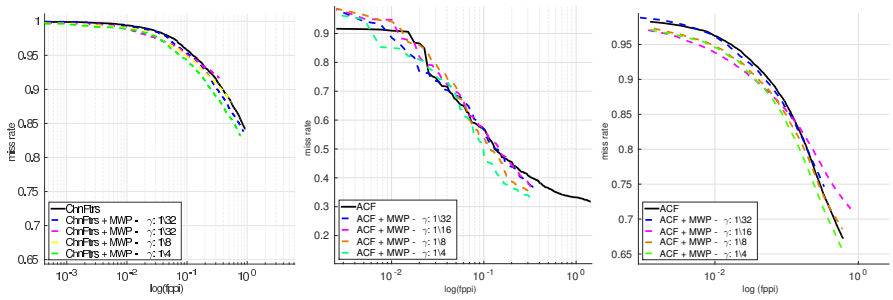
are present in this dataset. However, on the PETS S2.L1-view1 dataset, we obtain a significant reduction of the miss-rate, from 0.56 up to 0.41. In this case, in fact, MWP is able to discard wrong detections generated by DPM due to persons walking close each other. A little improvement is observed on the CAVIAR dataset for all γ values.



(a) TownCentre (b) PETS S2.L1-view1 (c) CAVIAR

Figure 4.5: Performance comparison on the TownCentre, PETS S2.L1-view1 and CAVIAR datasets using MWP with ChnFtrs detector

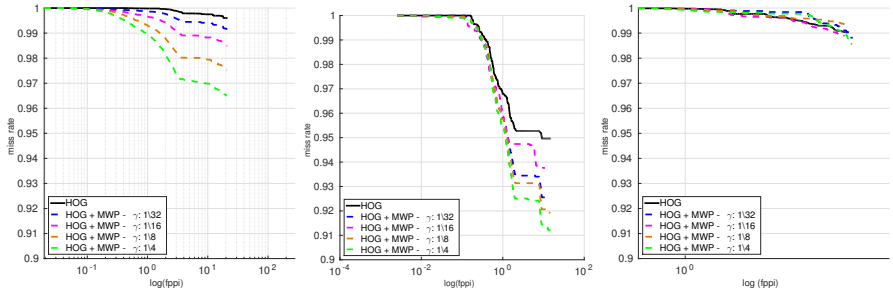
Fig 4.5 and Fig 4.6 show the performance considering the ChnFtrs and ACF person detectors, respectively. These two detectors are based on the same features and classifier. It is possible to observe a slight improvement in performance on both TownCentre and CAVIAR datasets using MWP.



(a) TownCentre (b) PETS S2.L1-view1 (c) CAVIAR

Figure 4.6: Performance comparison on the TownCentre, PETS S2.L1-view1 and CAVIAR datasets using MWP with ACF detector

However, a higher improvement can be appreciated on the PETS S2.L1-view1 dataset. This can be motivated by the fact that PETS S2.L1-view1 has a noisy background and MWP is able to discard regions that do not contains persons. Moreover, our strategy considerably improves the execution time of the ACF detector from about 20fps up to almost 60fps.



(a) TownCentre (b) PETS S2.L1-view1 (c) CAVIAR

Figure 4.7: Performance comparison on the TownCentre, PETS S2.L1-view1 and CAVIAR datasets using MWP with HOG detector

Fig 4.7 shows the ROC curves of the HOG person detector. The very low performance of this detector does not allow to appreciate the benefit introduced by our MWP. This is mainly due to the fact that HOG features have a really limited discriminative capability, resulting in a lot of missed

detections. However, the reduction of the total detections to be evaluated allows increasing detection efficiency from 2fps up to 14fps.

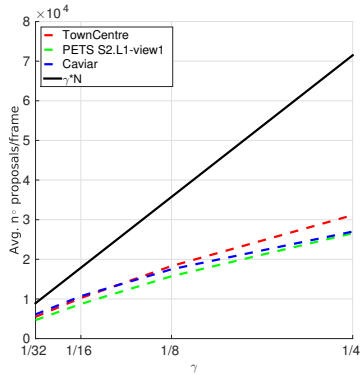


Figure 4.8: Detection windows generated by our MWP for the VeryFast detector on the test datasets. The maximum number of detection windows to be evaluated by the classifier is $N = 285944$.

Fig 4.8 shows the average number of detection windows generated with MWP for the three tested datasets and for different γ values. As reference, we also report the maximum number of windows N simply divided by γ. The observed gap is mainly due to the fact that MWP selects the window proposals considering the positions and scales that most likely contain a person according to the learned model, as well as the γ parameter, resulting in a very restricted set of final detection windows to be evaluated.

We have finally compared five state of the art sparse methods, namely SelectiveSearch [85], Objectness [51], EdgeBoxes [100], RandPrims [64] and Rath [75]. Tests were run on three publicly available datasets: TownCentre [16], PETS S2.L1-view1 [3] and CAVIAR [1].

As a baseline, we used the VeryFast detector, with the Soft Cascade classifier for the first octave trained on the Caltech Pedestrian dataset [38], and the Soft Cascade classifiers for the second and third octaves trained on the INRIA Person dataset [2], respectively. For the sparse methods, detection performance was evaluated running the VeryFast detector and applying the classifier exclusively to the sparse set of windows proposals of each method.

Figure 4.10 and Figure 4.11 show ROC curves on TownCentre and PETS S2.L1-view1 datasets. Table 4.5 shows the average number of frames per second for the proposal of the sparse set of windows and the average miss

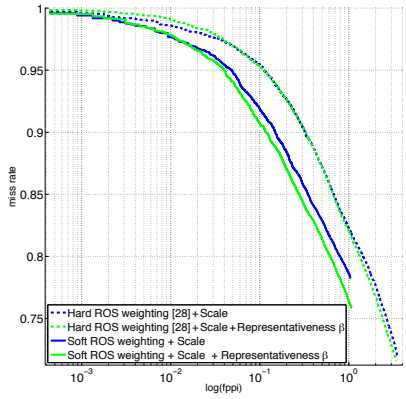


Figure 4.9: ROC curves of detection with sparse windows proposals (TownCentre dataset, $\gamma = 1/8$): effects of scale, representativeness β, Hard and Soft ROS.

	Scale	Scale + Representativeness
Hard ROS weighting [8]	0.95	0.95
Soft ROS weighting	0.92	0.91

Table 4.4: Miss-rate @10^{-1} FPPI of detection with sparse windows proposals (TownCentre dataset, $\gamma = 1/8$).

rate at 10^{-1} FPPI, respectively.

We can observe different behaviors of the methods with the two datasets. With the TownCentre, all the methods present a high number of false positives. The best miss rate is scored by MWP, EdgeBoxes and Objectness, with similar performance. However, both EdgeBoxes and Objectness have a much higher computational cost (see the number of frames per second required to obtain the sparse set in Table 4.5). With PETS S2.L1-view1, the difference in performance between MWP and the others is highly more apparent. In this case, the use of scale information in the GMM fitting of MWP reduces the number of false positives caused by the scene edgeness and cornerness. Similar speed-up as in TownCentre is observed in both the extraction of the sparse set and classification. Since both MWP and RandPrims allow to control the fraction γ of windows that is evaluated, effects of γ on performance is shown in Figure 4.12 and Figure 4.13 for these two

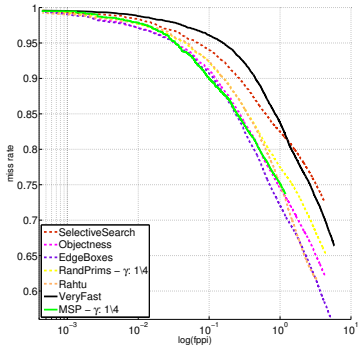


Figure 4.10: ROC curves of sparse windows methods on TownCentre.

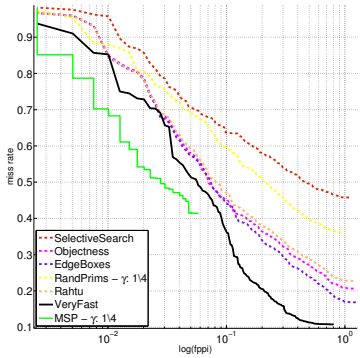


Figure 4.11: ROC curves of sparse windows methods on PETS S2.L1-view1.

methods. For MWP we can observe that reducing the number of windows in the sparse set produces only a slight miss rate decay (higher with PETS S2.L1-view1) but doubles its efficiency performance. This permits real-time operation of the method up to about 40 fps.

4.3 Conclusion

In this work we have proposed MWP as a new method that exploits a previously learnt scene model to provide sparse proposals of windows where targets should be detected most likely. The model of the scene is built using a Mixture of Gaussians whose components take into account position, scale and relative importance of detections in their Region of Support.

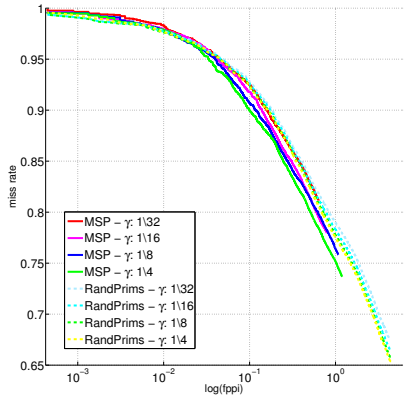


Figure 4.12: ROC curves of MWP and RandPrims on TownCentre.

Experimental results show that exploiting replacing sliding windows with our MWP in state of the art person detectors allows obtaining the lowest miss rate with a higher frame rate. This is mainly due to the ability of the method to select windows at the right scale of the target and to discard those regions that do not contains persons, so reducing the number of false positives.

Furthermore the set of filtered detections used to train our mixture model can be further exploited to train a specific classifier and further improve the detection accuracy, as also been done for scene-specific person detectors.

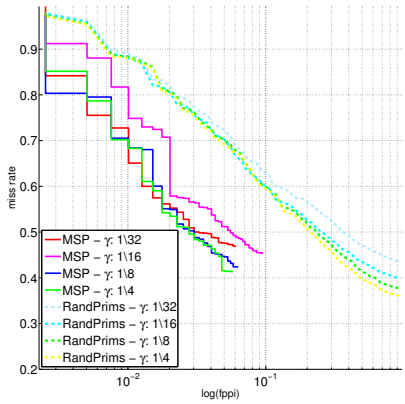


Figure 4.13: ROC curves of MWP and RandPrims on PETS S2.L1-view1.

Sparse method	TownCentre dataset		PETS S2.L1-view1 dataset	
	Sparse	Sparse & Classification	Sparse	Sparse & Classification
SelectiveSearch [85]	1.20	1.07 (0.94)	1.18	1.11 (0.63)
Objectness [51]	0.23	0.22 (0.91)	0.24	0.22 (0.45)
EdgeBoxes [100]	0.86	0.84 (0.91)	0.89	0.86 (0.44)
RandPrims [64] ($\gamma = 1/4$)	1.18	1.08 (0.92)	1.14	1.08 (0.60)
RandPrims [64] ($\gamma = 1/8$)	1.18	1.11, (0.93)	1.14	1.09 (0.60)
RandPrims [64] ($\gamma = 1/16$)	1.19	1.12(0.93)	1.14	1.09 (0.60)
RandPrims [64] ($\gamma = 1/32$)	1.20	1.14(0.93)	1.15	1.12(0.63)
Rautu [75]	0.17	0.16 (0.92)	0.20	0.18 (0.47)
MWP ($\gamma = 1/4$)	22.06	13.95 (0.90)	17.78	9.36 ($\sim$ 0.41)
MWP ($\gamma = 1/8$)	29.90	22.15 (0.91)	22.92	15.01 ($\sim$ 0.42)
MWP ($\gamma = 1/16$)	39.25	31.9 (0.92)	29.29	22 ($\sim$ 0.45)
MWP ($\gamma = 1/32$)	48.92	40.65 (0.92)	36.71	30.47 ($\sim$ 0.47)

Table 4.5: Comparative performance of the sparse methods. Number of frames per second needed for sparse windows proposal and sparse windows proposal with classification. Miss rates at 10^{-1} FPPI are also shown in parenthesis.

Chapter 5

User Interest Profiling Using Tracking-free Coarse Gaze Estimation

Understanding where people attention focuses is a challenging and extremely valuable task that can be solved using computer vision technologies. In this Chapter we address this problem on surveillance-like scenarios, where head and body imagery are usually low resolution. We propose a method to profile the attention of people moving in a known space. We exploit coarse gaze estimation and a novel model based on optical flow to improve attention prediction without the need of a tracker. Removing the tracker dependency makes the method applicable also on highly crowded scenarios. The proposed method is able to obtain comparable performance with respect to state of the art solutions in terms of Mean Average Angular Error (MAAE) on the TownCentre dataset. We also test our approach on the publicly available MuseumVisitors dataset showing an improvement both in terms of MAAE and in terms of accuracy in the estimation of visitors' profile.

5.1 Introduction

Understanding what objects are looked at and for how long is also of great interest for retail companies that may want to obtain a large dataset of customer behavior. This is often solved by tracking all the persons in the scene and consequently generating heat images, registered with the shop maps, that indicate customer persistence. Although, even if the scene strongly constrains people position, such as in a supermarket aisles, there is a lot of ambiguity if we consider just the position. If we are willing to detect which products draw people attention in a shop, gaze estimation is the only option.

A slightly different but complementary task is profiling the interests of a single person in a given environment. In this case, instead of accumulating a global statistic from all persons behavior, a single profile is sought.

Passive profiling finds several interesting applications in the cultural heritage scenario [53]. For example, user profiling can help solving many issues Museums struggle to cope with; like personalizing content for visitors. Personalization should both increase engagement and satisfaction creating a dedicated view of museum collections and suggesting novel cultural paths to explore. Moreover a recommender system may, also building from previously watched people behaviors, help in planning further tours towards different cultural venues, places of interest or museums.

To this end both person gaze and position in the scene are very relevant to understand the attention; how far an object is from the person could not be a sufficient hint. We argue that understanding which objects are in the person's field of view is crucial for a correct attention estimation.

We propose a method for coarse gaze estimation that can be exploited for video surveillance, for the analysis of social behavior interaction and for attention profiling. Our solution exploits frame-to-frame motion information and therefore does not need to track every person in the scene, as in [16], or perform complex and computationally onerous global optimization requiring the knowledge of the entire person trajectory.

5.2 Stateless coarse gaze estimation

In this section we first summarize how to learn a model that is able to estimate at runtime coarse head and body poses. Then we introduce a motion model to improve the coarse gaze estimation for moving persons.

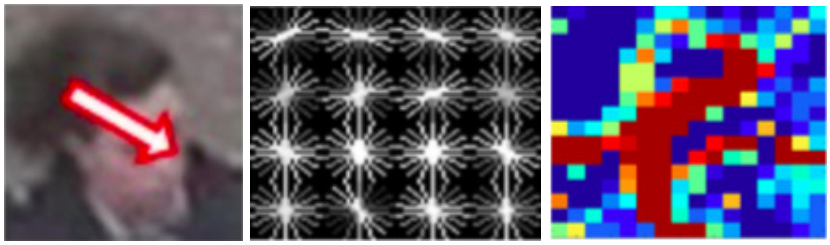


Figure 5.1: Feature extracted from a sample head.

To detect person in the scene we use the detector from [42] that is able to segment both the body and the head of the detected person.

5.2.1 Head and body pose estimation

We build upon the solution proposed in [15,17] in order to coarsely estimate the head and body orientations.

For the head visual representation, we resize each patch to a standard resolution of 128x128 pixels from which we extract the Histogram of Oriented Gradients (HOG). Then we resize the same patch to a resolution of 16x16 pixels and extract both the intensity of the gradients and the RGB colors. The final head descriptor is obtained as early fusion of these three distinct features and has a dimension of 1600 bins: 576 bins for the HOG feature, 256 (16x16) bins for the intensity of the gradient and 768 (16x16x3) bins for the RGB color channels. A sample of the feature extraction process is reported in Fig. 5.1.

We use *random ferns* [18], as in [15], to train our model, and we will refer to it as *Head-ferns*. The *fern* differ from the standard decision trees since the same set of branch-test is applied to each image regardless of the previous test results. We quantize all the possible head orientations (from 0 to 360 degrees) in 16 classes.

Estimating the orientation of the head can be really difficult due to the limited resolution at which a head is observed in typical surveillance footage and also because of missing information about the context in which the head is acquired. Indeed, the class with the maximum score given by the Head-ferns does not always represent the correct orientation. It could happen that there are two or more modes and in this case choosing the orientation class with the highest score can lead to a wrong decision. For this reason, we would

like to refine the initial estimation given by our Head-ferns by exploiting the whole body orientation, as also proposed in [25].

As for the head, we train *random ferns* using as input a set of features extracted from the whole body image of a person. In particular, we extract the same features of the head but we resize the body patches to a standard resolution of 384x128 pixels for the HOG and 48x16 pixels for the intensity of the gradients and the RGB colors (we keep an aspect ratio of 3:1). For the random-ferns we quantize the possible orientation in 8 classes; we will refer to this model as *Body-ferns* from now on.

We finally concatenate the output of both the Head- and Body-ferns predictors to form a new set of features and train a SVM classifier with a RBF kernel. We cross-validate the regularization parameter C and estimate σ as the average distance between training features.

5.2.2 Motion model

The use of the head and body orientations may not always be sufficient to correctly discriminate the gaze of a person. This is mainly motivated by two reasons: 1) low resolution patches can be too ambiguous to be discriminated by the classifier; 2) for body patches it is really difficult to discriminate between a person seen frontal (0 degrees) or rear (180 degrees). For this reason some solutions have been proposed in literature that exploit tracking information to constrain the gaze of a person towards its direction. This information can be particularly useful for moving people. However, tracking all the persons in a scene is computationally onerous and prone to failure due to drift issue.

For this reasons we introduce a motion feature in our gaze representation. We believe that just the motion of a person can instantly disambiguate such situations. We use the technique from [63] to extract the optical flow from two consecutive frames at time I_{t-1} and I_t. We discard all those pixels with a motion below a given threshold τ and then compute the optical flow orientation for the remaining pixels. For each bounding box detected in the image I_t we compute the histogram of orientations weighted according to an Epanechnikov kernel. We quantize the possible orientation in 8 classes. We will refer to this feature as *Histogram of Oriented Optical Flow (HOOF)*. Fig. 5.2 shows the HOOF extraction process.

The use of this feature allows us to keep our solution stateless while granting a lower computational cost with respect to solutions based on tracking

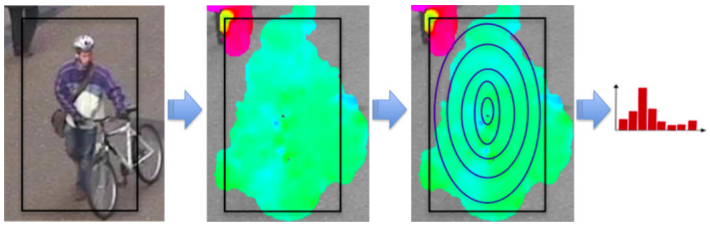


Figure 5.2: Motion feature extracted from a person detection.

or global optimization.

The final model is learned using as features the concatenation of the predictions from the Head- and Body-ferns and the HOOF motion feature. As in the case of the concatenation of head and body orientation prediction we learn an SVM to predict the final gaze.

5.3 User Profiling

Our goal is to identify for each person the interest towards the surrounding environment. For this purpose the estimation of the gaze of a person can be used to determine an area of the scene that represent, with high probability, the subject of user's attention. To this end the coarse gaze estimated as in Sect. 5.2 can be exploited to profile user interests in a scene and give him more details about its preferences.

In order to be able to understand where the person is looking to or at what is looking at in the observed scene we need to: 1) map the position and gaze of a person on the ground plane; 2) compensate the projection of the gaze [77] with respect to the real world reference system. To this end we first estimate the camera matrix $\mathbf{H}$ using the intrinsic and extrinsic camera parameters. Then it is possibile to estimate the compensation needed for the gaze as:

$$\boldsymbol{\theta} = \arccos\left(\frac{\mathbf{Hv} - \mathbf{Hx}}{||\mathbf{Hv} - \mathbf{Hx}||} \cdot \mathbf{i}\right) \tag{5.1}$$

where $\mathbf{x}$ is the position of the target in the image plane and $\mathbf{v}$ is the vanishing point, see Fig. 5.3.

Once both position and the gaze are projected it is possible to exploit these information to profile the interests towards the environment for each

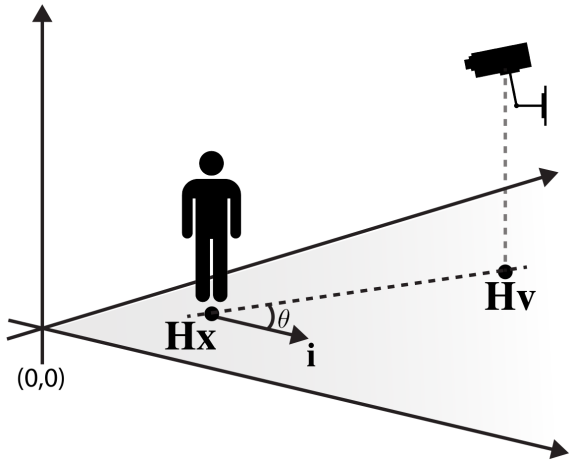


Figure 5.3: Visual representation of how the compensation angle θ is computed.

person and, vice versa, understand which objects (e.g. artworks in a museum) of the scene are more attractive. For each object position $\mathbf{Hx}_k$ and each person position $\mathbf{Hx}_i$ on the ground plane we define:

$$d_{ik}(\alpha) = \alpha \frac{||\mathbf{p}_{ik}||}{M} + (1-\alpha) \arccos\left(\mathbf{g}_i(\boldsymbol{\theta}) \cdot \frac{\mathbf{p}_{ik}}{||\mathbf{p}_{ik}||}\right) \pi^{-1} \tag{5.2}$$

where

$$\mathbf{p}_{ik} = \mathbf{Hx}_k - \mathbf{Hx}_i \tag{5.3}$$

being $\mathbf{g}_i(\boldsymbol{\theta})$ the person's gaze projected on the ground plane through $\mathbf{H}$ and corrected with the angle $\boldsymbol{\theta}$, M the maximum distance an artwork can have from a visitor in the room and α a factor that weighs the combination of the distance between the person i and the object k with the person's gaze.

The artwork k^* to be assigned to the person's profile is selected using:

$$k^* = \arg\min_k d_{ik}(\alpha). \tag{5.4}$$

Note that if $\alpha = 1$ we obtain the naif model associating people to artworks based only on the position on the ground plane.

5.4 Experiments

In this section we report a set of experiments to assess the performance of our solution for coarse gaze estimation in comparison with state of the art methods. Then we show how estimating the interest of a person through both position and gaze improves with respect to using just the position of a person in the scene.

5.4.1 Datasets and experimental details

Tests are conducted on two different datasets, TownCentre [16] and MuseumVisitors [9]. The TownCentre dataset is a outdoor surveillance video composed of 4500 frames with high scale variations for each person, occlusions, and false positives in the scene. We randomly split the set in 218 persons for the training and 57 persons for the test.

MuseumVisitors is a challenging dataset recorded at National Museum of Bargello in Florence, composed of three sequences acquired with three IP cameras at a resolution of 1280×800 pixels. This dataset is specifically designed for group detection, occlusion handling, tracking, re-identification and behavior analysis. On MuseumVisitors we adopted the leave-one-out strategy to evaluate our solution, so one person detection is used as test while the other detections are used for training. The final accuracy is obtained by averaging over all the results.

The ferns for the head orientation have been trained using the BMVC2009 dataset [15], that contains 1477 cropped head taken from different viewpoints, with resolution from 10×10 pixels to 128×128 pixels. While the ferns for the body have been trained on the TUD dataset [25], considering 7657 body patches extracted from 4732 frames, with resolution from 79×26 pixels to 310×102 pixels. For both Head-ferns and Body-ferns, the number and the size of each fern have been chosen experimentally through a phase of preliminary validation. In particular, we use 200 ferns each with a size of 10, respectively.

5.4.2 Gaze estimation evaluation

In this section we describe the improvements introduced by using different features with the proposed strategy. In particular, we analyse the performance between exploiting Head (H) and Body (B) ferns predictors, and

Histogram of Oriented Optical Flow (O) alone and their combinations. The results are reported in terms of Mean Absolute Angular Error (MAAE) computed between the estimated gazes $\{g_i\}$ and the ground truth $\{G_i\}$ on the image plane:

$$MAAE = \frac{1}{N}\sum_{i=1}^{N} \min\{|g_i - G_i|, |g_i - G_i \pm 360^\circ|\}.$$

Table 5.1 shows the performance of our strategy compared with Benfold et al. [17] and Chen et al. [25] methods on the TownCentre dataset. We specify the characteristics of each strategy in terms of using Head or Body gaze estimation, motion and tracking. We consider a method using *motion* if it exploits as cue the information computed from two adjacent frames such as the walking direction or the optical flow. We consider a method using *tracking* if it uses the information from multiple frames to estimate a single gaze. This can be done in a causal and non-causal manner, in this latter case performing a global optimization.

On TownCentre, our strategy with only the motion feature obtains comparable result with respect to the other methods. This is mainly due to the fact that in the TownCentre dataset the person walks in the street with gaze mainly oriented towards the motion direction. Our best with 22° of MAAE is obtained with the full features combination. Although, Chen et al. [25] reach the lowest MAAE, that is 18°, the strong limitation of this method is the use of tracking information to extract the gaze, which reduces the applicability of the method in real scenarios where occlusions and crowd are present.

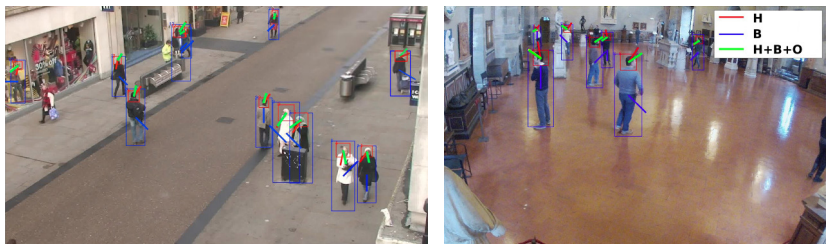
In Table 5.2 we report the performance obtained on the MuseumVisitors, considering only the persons with occlusion area lower than 20%. In particular, we evaluate 1400 persons in Camera 1, 166 persons in Camera 2 and 1192 persons in Camera 3. The gap in performance varying the features is notable. Using only Optical Flow produces the worst results on all cameras, with gaze errors over 40°. The Head feature reduces the error in the cameras 1 and 3 with respect to Body and Optical features. A larger improvement is achieved by combining Head and Body, that drops the gaze error. Best results are obtained exploiting the combination of all features with an error lower than 30° on all cameras. This is mainly due to the fact that the direction extracted from the motion of each person limits the range of feasible gazes in our method, improving the accuracy. In Fig. 5.4 we show

Strategy	MAAE	Head Gaze	Body Gaze	Motion	Tracking
Benfold [17]	26°	✓	✗	✗	✓
Benfold [17]	26°	✓	✗	✓	✓
Chen [25]	45°	✓	✗	✗	✓
Chen [25]	28°	✓	✓	✓	✗
Chen [25]	18°	✓	✓	✓	✓
Our (O)	26°	✗	✗	✓	✗
Our (H)	42°	✓	✗	✗	✗
Our (B)	45°	✗	✓	✗	✗
Our (H+B)	42°	✓	✓	✗	✗
Our (H+B+O)	22°	✓	✓	✓	✗

Table 5.1: Mean Absolute Angular Error of the proposed strategy in comparison with state-of-the-art on the TownCentre dataset.

Feat. Combination	Camera 1	Camera 2	Camera 3
O	46°	47°	51°
H	34°	35°	34°
B	35°	30°	43°
H+B	28°	26°	32°
H+B+O	**26°**	**22°**	**30°**

Table 5.2: Mean Absolute Angular Error on the MuseumVisitors dataset with the proposed method (for different features combination).



(a) *TownCentre* (b) *MuseumVisitors: Camera 1*

Figure 5.4: Example of persons' gaze estimated with the proposed strategy in TownCentre (a) and MuseumVisitors (b).

the gaze extracted with the proposed strategy in one frame of Camera 1 of the MuseumVisitors dataset and on a frame from the TownCentre dataset. MuseumVisitors is a more challenging dataset for gaze estimation as it can be seen gaze can be hardly inferred by people motion alone, while on TownCentre gaze is almost parallel to the walking direction. Indeed, our method only using optical flow (O), as is shown in Table 5.2, is much worse than in Table 5.1.

5.4.3 Profiling evaluation

In this section we report the accuracy of user profiling on MuseumVisitors. For the test we considered 10 artworks inside the Donatello's Hall, as shown in Fig 5.5. An interesting annotation that is provided with this dataset is the association, for each frame, of visitors to artworks. The ground truth also specifies if no relevant object is observed by a person. We measure the accuracy of correct visitor-artwork association. If $d_{ik}(\alpha) > 0.2$ we do not associate a visitor to any artwork.

In Table 5.3 we report the accuracy of the computed profiles, considering the geometrical distance alone ($\alpha = 1$) and the combination of distance and gaze ($\alpha < 1$). In the last case, we report only the best results obtained with $\alpha = 0.75$. In general, the performance improves using the distance and gaze together, reaching the highest accuracy with the combination of all features. Some sample of correct and wrong association for different setup of our method are shown in Fig 5.6.

Finally, in Fig 5.7 we show, for each camera and over all the cameras,

Score function	Camera 1	Camera 2	Camera 3
Geom. distance: $d_{ik}(1)$	88%	69%	84%
$d_{ik}(0.75)$ + Feat. O	87%	60%	82%
$d_{ik}(0.75)$ + Feat. H	91%	68%	86%
$d_{ik}(0.75)$ + Feat. B	90%	69%	86%
$d_{ik}(0.75)$ + Feat. H+B	91%	73%	86%
$d_{ik}(0.75)$ + Feat. H+B+O	**93**%	**75**%	**86**%

Table 5.3: Accuracy of the profiles of interest varying the features combination of the proposed method.

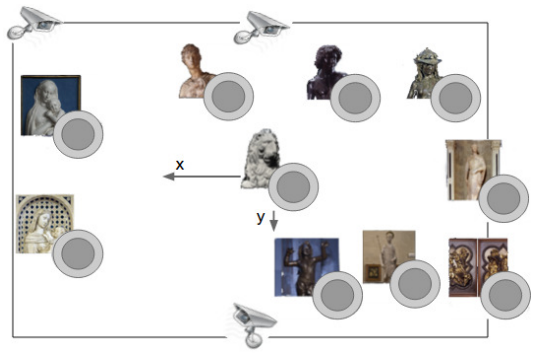
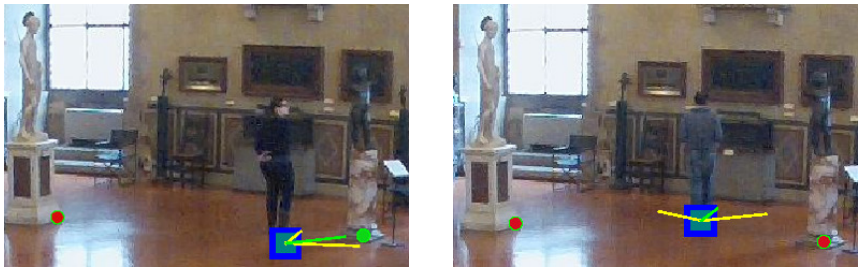


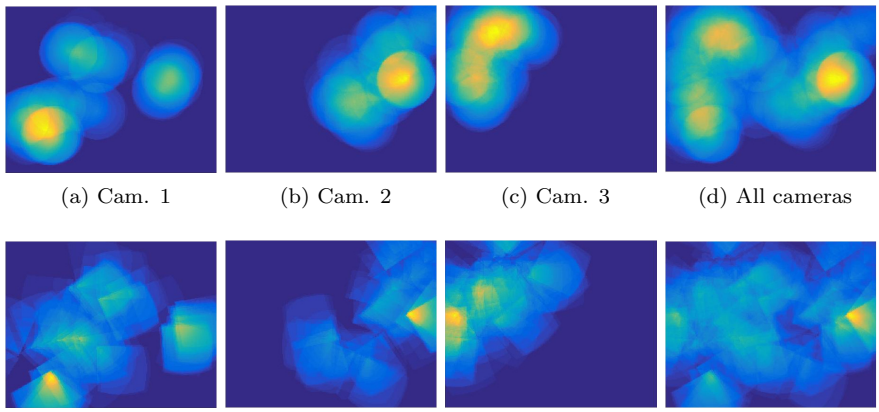
Figure 5.5: Artworks location inside the Donatello's Hall.

the heatmap obtained using the position of the persons in the scene and the heatmap obtained using both the position and the gaze. It can ne noted that the gaze heatmap is more informative. Indeed if we compare the maps from camera 3, the position heatmap (c) estimates a lot of energy in the top left corner of the room, while for the gaze map (g) the area is not receiving any interest. This is a more realistic prediction since the corner does not contain relevant artworks and the two artworks on the left side are minor works, with less historical and artistic relevance with respect to the Donatello's sculptures on the other side of the room.



(a) (b)

Figure 5.6: Anecdotal evidence of our approach: (a) correct association by gaze or position; (b) wrong artwork association using position while no artwork is actually looked at.



(a) Cam. 1 (b) Cam. 2 (c) Cam. 3 (d) All cameras

(e) Cam. 1 (f) Cam. 2 (g) Cam. 3 (h) All cameras

Figure 5.7: Heatmaps of the profiles of interest in the Donatello's Hall computed considering the feet position on the ground plane (first row) or the combination between feet position and gaze (second row).

Chapter 6

The Mnemosyne System: delivering personalized information to Museum visitors

The amount of multimedia data museums gather in their databases is growing fast, while the capacity to display more information to visitors is limited. Such information often targets the interests of average visitors instead of the whole spectrum of different interests each individual visitor could have. In this Chapter we describe the Mnemosyne system that addresses these issues through a new multimedia museum experience. Mnemosyne builds a user profile for each visitor used to drive an interactive table to personalize the multimedia content delivery of the available resources.

6.1 Introduction

Artworks displayed in popular museums range from few thousands to hundreds of thousands. Visitors are challenged with a huge amount of information. Every single piece of exhibited art comes with many layers of additional contextual information that people must filter according to their own interest, need and time available. Museum directors usually design hall layouts and art descriptions to target an "average visitor". It would not be possible to provide the full history, contextual information, author biography, related

artworks and sites, for every work of art in display. This information must be delivered in an organized manner, allowing each individual to sift through the available data following her own inclinations and interests.

Moreover, some visitors may prefer to maximize their visual and sensory experience, discarding historical and technical artistic details. Even the most culturally aware may better like the possibility to defer the deepening of their knowledge of seen art at a later time.

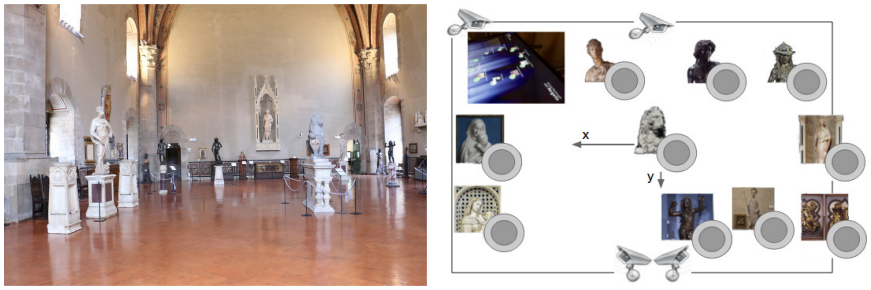
Museums should exploit modern technology to improve visitor experience and engagement. The main goal of such approaches should be that of intercepting user interest and expand the amount of information accordingly. Many efforts in the past exploited augmented reality applications. The main drawback of this interaction paradigm is the intrusiveness, since it requires the user to employ a hand-held device every time she feels the need to get more information on a work of art. This approach although potentially leading to visually pleasant and informative views of the real world, completely disrupts the visitor attention, and in our opinion the quality of her experience.

Another way of gathering information on user interest is through profiling. Profiling means to associate to every visitor a probability distribution over the artworks, representing her interest.

Passive profiling is the task of gathering user attention measurements. Since a direct measurement of one's attention is not physically possible one must rely on cues that proxy the actual attention. A strong cue is the physical proximity to an artwork. How far a person is from a certain physical point can be measured using several strategies. Wireless signals have been used in the past for localization, but state-of-the art technologies may have errors even up to two meters. Moreover in crowded environments signals are highly disturbed.

Accurate results may be obtained coupling a mobile application with BLE tags. This approach is unfortunately, also intrusive, needing to deploy tags on every artwork and the user to install some app to keep track of tag activations.

In this work we propose to use passive user profiling through computer vision techniques. We couple the computer vision system with a mobile app and an interactive table for the deferred information delivery. Through computer vision techniques we can acquire very precise measurements of people location using calibrated cameras. Moreover, head visual features



(a) Overview of the Hall (b) Artworks monitored from the system

Figure 6.1: Mnemosyne installation at the Donatello's Hall.

can be exploited to estimate a coarse gaze, improving the cues for attention further.

6.2 Mnemosyne System

Mnemosyne is a working installation opened in February 2015 at the National Museum of Bargello, in the Donatello's Hall (see Fig. 6.1(a)). Inside this hall there are more than 70 artworks, most of them realized by Donatello. We consider ten artworks of interest, as shown in Fig. 6.1(b).

Based on passive observation of the tourists during the visit, the system extracts a profile of interests. This is obtained by analyzing, in real-time, the video streams of four fixed cameras opportunely positioned in the Donatello's hall. For each camera the system localizes the tourists and build local profiles considering all the artworks observed by each visitor. The profiles of the same person coming from different cameras are then merged to obtain an unique global profile. At the end of the visit, the tourist can interact with a user interface that shows personalized contents according to the estimated global profile and exploit a mobile application to download a summary of his visit with additional multimedia contents.

The execution pipeline is composed by the following stages:

- Visitor Detection: localization of persons in the frames of each camera.
- Visitor Description: extraction of a visual person descriptors from each detection to capture the visual characteristics of each visitor.

(a) Camera 1 (b) Camera 2 (c) Camera 3

Figure 6.2: Different challenges for pedestrian detection in real scenarios: scales variation of the target (red and green box), false positives in the image (violet box), strong occlusion and crowd of the people (blue box).

- Visitor Local Modelling: build of a local profile for each visitor based on the observed artworks.
- Visitor Global Modelling: merging of local profiles between the four cameras.

In the following we will expound each stage of the Mnemosyne pipeline.

6.2.1 Visitor Detection

Pedestrian detection in a real museum is a very difficult task due to many challenges that must be addressed, as shown in Fig. 6.2. The first issue to be considered is the high scale variation of pedestrians in the image due to scene perspective. Usually, multiple scales of the same image are considered (pyramid of the image) and a sliding windows is evaluated by the classifier at each location and scale of the image.

To overcome these issues a model of the scene is built that is based on the expected scale of each visitor in the image. The model is learnt in an unsupervised way, considering as training set the output of a pedestrian detector. At test time, according to the scene model, only the detection windows at the correct height are generated and evaluated by the classifier, producing an improvement in terms of both accuracy and speed. The scene model is defined as a linear combination of the visitors positions $\{x, y\}$ in the image:

$$f(x, y, \boldsymbol{\gamma}) = \boldsymbol{\gamma} \cdot [1 \ x \ y]^T, \qquad \boldsymbol{\gamma} \in \mathbb{R}^3 \tag{6.1}$$

where $\boldsymbol{\gamma}$ are the plane parameters. To estimate $\boldsymbol{\gamma}$ we consider a set of positive detection windows $\{d^{(i)} = [d_x^{(i)}, d_y^{(i)}, d_w^{(i)}, d_h^{(i)}]\}$ extracted from the same

camera, where $d_x^{(i)}$ and $d_y^{(i)}$ represent the coordinates of the window in the image, $d_w^{(i)}$ and $d_h^{(i)}$ are respectively the width and height of the detection window. For each detection $d^{(i)}$, given the classification score $d_s^{(i)}$ of the detector, we estimate its relevance, $\pi^{(i)}$, by considering the cardinality of its ROS:

$$\pi^{(i)} = |\{d^{(r)} |\, d_s^{(r)} \leq d_s^{(i)} \text{ and } \frac{d^{(r)} \cap d^{(i)}}{d^{(r)} \cup d^{(i)}} \geq \tau_{voc}\}|$$

where $\frac{d^{(r)} \cap d^{(i)}}{d^{(r)} \cup d^{(i)}}$ is the VOC score computed as the intersection over union of the bounding boxes $d^{(r)}$ and $d^{(i)}$ while τ_{voc} is a threshold. A weighting factor $w^{(i)}$ is then assigned to each $d^{(i)}$ based on its relevance and classification score:

$$w^{(i)} = \pi^{(i)} \cdot d_s^{(i)}, \qquad d_s^{(i)}, \pi^{(i)} \in \mathbb{R}$$

The best configuration for the plane parameters $\boldsymbol{\gamma}$ is finally obtained by solving the following weighted least square problem:

$$\boldsymbol{\gamma} = \arg\min_{\boldsymbol{\gamma}} \sum_i w^{(i)} [d_h^{(i)} - f(d_x^{(i)}, d_y^{(i)}, \boldsymbol{\gamma})]^2 = (\mathbf{D}_{x,y}^T \mathbf{W} \mathbf{D}_{x,y})^{-1} \mathbf{D}_{x,y}^T \mathbf{W} \mathbf{D}_h$$

6.2.2 Visitor Description

We describe the visual appearance of each person detected in the scene through a descriptor $d_a^{(i)}$ composed of both color and texture features [61]. In particular, color information are encoded as histograms in the HS, RGB and Lab color spaces while texture information is described through Histogram of Oriented Gradients (HOG) [30].

Together with the visual description we collect over time both spatial and temporal information for each detection. Spatial information $d_g^{(i)}$ are represented through the absolute position of the person on the ground plane, while the synchronized timestamp $d_t^{(i)}$ represents the temporal information.

6.2.3 Visitor Local Modelling

The path covered by each visitor in a camera is described in terms of small groups of detections clustered together, namely tracklets. Tracklets are created exploiting the visual description jointly with spatial and temporal information associated with each detection.

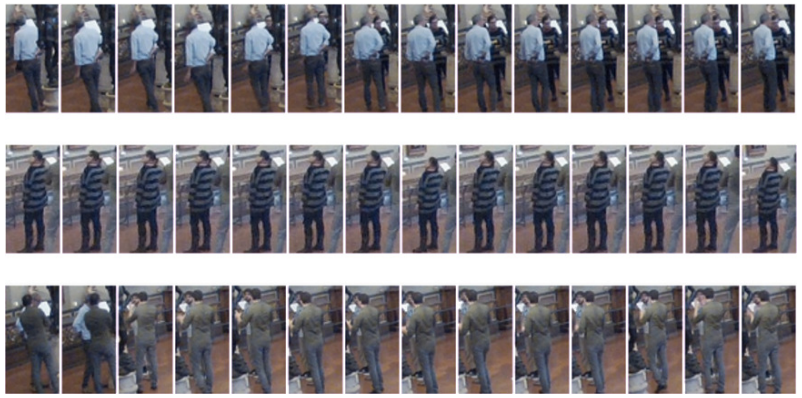


Figure 6.3: Example of tracklets generated for three different persons.

To assign a new detection $d^{(i)}$ to an existing tracklet $m^{(j)}$ the similarity is computed as follows:

$$\delta(d^{(i)}, m^{(j)}) = exp\Big\{ - \Big((1-\alpha-\beta)||d_a^{(i)} - m_a^{(j)}||_2 + \alpha\frac{||d_g^{(i)} - m_g^{(j)}||_2}{w_g} + \tag{6.2}$$

$$+ \beta\frac{||d_t^{(i)} - m_t^{(j)}||_2}{w_t}\Big)\Big\} \tag{6.3}$$

where $m_a^{(j)}$ is the appearance descriptor of the detection in the median position of the j-th tracklet, while $m_g^{(j)}$ and $m_t^{(j)}$ are respectively the spatial and temporal information of the last detection in $m^{(j)}$. The normalization factors w_g and w_t define respectively the spatial and temporal intervals of observation, while α and β control the contribution of each component to the distance score. The detection $d^{(i)}$ will be associated to an existing tracklet only if the similarity score $\delta(d^{(i)}, m^{(j)})$ is greater than a fixed threshold and according to:

$$j^* = \arg\max_j \ \delta(d^{(i)}, m^{(j)}) \tag{6.4}$$

A new tracklet will be created every time a detection cannot be associated to any of the existing tracklet. In figure 6.3 are shown three tracklets obtained considering the eq. 6.4.

The local profile $p^{(j)}$ associated to $m^{(j)}$ is computed considering the distance between each detection in the tracklet and the gaze of the person.

In fact, building a profile of a visitor without considering orientation information can be misleading. In many situation a person may stand close to an artwork but look in an opposite direction. To this end, we propose to include coarse gaze information in the local profiling model. In particular, we build upon our previously proposed solution (Chapter 5) which does not require a tracker to obtain reliable gazes, but exploits optical flow as a cue for incorporating motion information. Our final coarse gaze integrates head, body and motion orientations.

In order to be able to understand where the person is looking to or at what is looking at in the observed scene we need to: 1) map the position and gaze of a person on the ground plane; 2) compensate the projection of the gaze [77] with respect to the real world reference system. To this end we first estimate the camera matrix $\mathbf{G}$ using the intrinsic and extrinsic camera parameters. Then it is possibile to estimate the compensation needed for the gaze as:

$$\boldsymbol{\theta} = \arccos\left(\frac{\mathbf{G}\cdot\mathbf{v} - \mathbf{G}\cdot[dx^{(i)}, dy^{(i)}, 1]'}{||\mathbf{G}\cdot\mathbf{v} - \mathbf{G}\cdot[dx^{(i)}, dy^{(i)}, 1]'||}\cdot\mathbf{i}\right) \tag{6.5}$$

where $\mathbf{v}$ is the vanishing point.
Once both position and the gaze are projected it is possible to exploit these information to profile the interests towards the environment for each person and, vice versa, understand which objects (e.g. artworks in a museum) of the scene are more attractive. For each artwork position and each person position on the ground plane we compute the follows distance:

$$d_k^{(i)}(\alpha) = \alpha\frac{||\mathbf{p}_{ik}||}{M} + (1-\alpha)\arccos\left(\mathbf{g}_i(\boldsymbol{\theta})\cdot\frac{\mathbf{p}_{ik}}{||\mathbf{p}_{ik}||}\right)\pi^{-1} \tag{6.6}$$

where

$$\mathbf{p}_{ik} = \mathbf{H}\mathbf{x}_k - \mathbf{H}\mathbf{x}_i \tag{6.7}$$

being $\mathbf{g}_i(\boldsymbol{\theta})$ the person's gaze projected on the ground plane through $\mathbf{G}$ and corrected with the angle $\boldsymbol{\theta}$, M the maximum distance an artwork can have from a visitor in the hall, and α that weighs the combination of the distance between the person i and the artwork k with the person's gaze. Each detection $d^{(i)}$ of the tracklet $m^{(j)}$ contributes to the k-th bin of the local profile histogram $p^{(j)}$ as follows:

$$p^{(j)} = \{p_k^{(j)}\}_{k=1}^{K}, \qquad \text{where} \quad p_k^{(j)} = \sum_{\forall d^{(i)}\in m^{(j)}} e^{-\frac{1}{2}\left(d_k^{(i)}(\alpha)\right)^2} \tag{6.8}$$

6.2.4 Visitor Global Modelling

When a visitor approaches the tabletop, its detection d is used to retrieve the most similar tracklets $m^{(j)}$ from each camera and the relative local profiles $p^{(j)}$. These profiles are fused together in order to obtain the global profile p of the visitor considering each artwork k, as follows:

$$p = \{p_k\}_{k=1}^{K}, \qquad \text{where} \quad p_k = \sum_{\forall m^{(j)} \,|\, ||m_a^{(j)} - d_a||_2 < T} \frac{p_k^{(j)}}{1 + e^{-6\left(\frac{1}{||m_a^{(j)} - d_a||_2}\right)}}$$

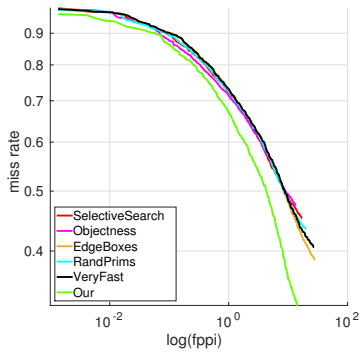
where T is a threshold that avoids considering those tracklets with a high distance, in terms of appearance, with respect to appearance extracted from the visitor detection d.

6.3 Experiments

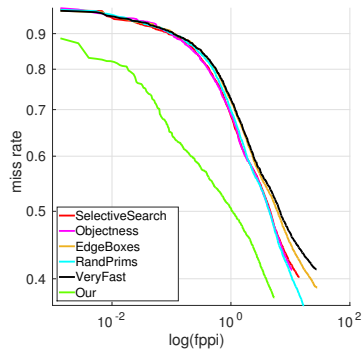
In this section, we report a set of experiments to assess the performance of each stage of the Mnemosyne pipelinem, and the final accuracy of the user profiling. Tests are performed on a dataset composed of four synchronized sequences, one for each camera in the Donatello's hall, recorded during a real visiting scenario. Each sequence is composed about 2000 frames and 67 observed visitors. For each person across each camera we manually labelled the identity, the bounding box, the visible region of the bounding box and the associated artwork if he is observing any.

6.3.1 Visitor Detection

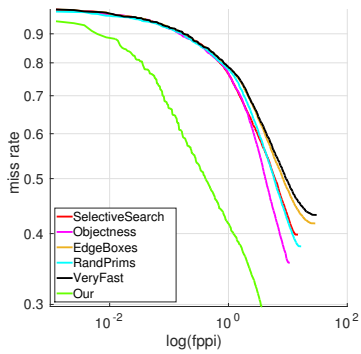
We evaluated the performance of the our strategy with state of the art methods to obtain sparse detection windows proposals, using the VeryFast detector [14] to evaluate each detection window. Experiments were conducted considering all the bounding boxes from the ground truth or considering a subset of these composed by all the bounding boxes with a minimum height of 100 pixels and a percentage of occlusion lower than 50%, namely *reasonable*. We have analyzed five state of the art sparse methods: SelectiveSearch [85], Objectness [51], EdgeBoxes [100] and RandPrims [64]. Results are shown in Fig. 6.4. We can observe different behaviors of the methods in the four cameras. In Camera 1, we obtain similar miss rate performance as the other



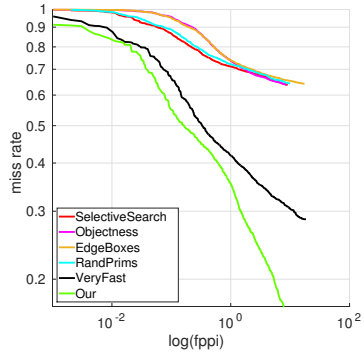
(a) Camera 1



(b) Camera 2



(c) Camera 3



(d) Camera 4

Figure 6.4: ROC curves of sparse windows methods for each camera.

	Cam. 1				Cam. 2				Cam. 3				Cam. 4			
Rank	1	10	20	40	1	10	20	40	1	10	20	40	1	10	20	40
	85.77	97.82	98.91	99.47	31.01	67.57	78.49	85.04	26.36	63.73	75.86	83.14	27.13	64.38	72.54	78.33

Table 6.1: Performance of the visitor re-identification for each camera, considering the single-vs-all scenario at different ranks.

strategies, with a slight improvement in terms of accuracy. With the Camera 2, all the tested methods present a high number of false positives at $10^{-1}FPPI$. The lowest miss rate is scored by our strategy. With Camera 3, the difference in performance between the proposed solution and the others is highly more apparent. In this case, the use of the proposed plane of the scales reduces the number of false positives caused by the scene complexity. In Camera 4, we obtain similar performance as the VeryFast, but with a much lower computational cost with respect to the dense method. We can observe that reducing the number of windows to be evaluated does not produce a loss in the accuracy, but doubles its efficiency performance. This permits to run out person detection in real-time, at about 40 fps.

6.3.2 Visitor Re-identification

We evaluated the re-identification performance between cameras considering the person descriptor [60] and a simple nearest neighbor classification. Tests were conducted following a single-vs-all (SvsAll) scenario [54]. We consider as probe images all those person' detected in front of the interactive table (in Camera 1), the rest of the detections from all the four cameras are used as gallery. Identity knowledge is not exploited for both probe and gallery. Results are reported in Table 6.1. We can observe a difference in terms of accuracy between the Camera 1 and the rest of the cameras. This result can be explained considering that the probe and the gallery set come from the same camera, sharing identical conditions of illumination and prospective of the images. In this case, best performance reach over 85% of accuracy at rank 1. Likewise, similar results are obtained with the Cameras 2 and 3 due to the same complexity and prospective in both the cameras. Totally different results in Camera 4, where the reduced brightness in the images and the strong presence of false positive in the scene compromise the identification of the persons.

		Cam. 1		Cam. 2		Cam. 3		Cam. 4	
		#Traces	Acc.	#Traces	Acc.	#Traces	Acc.	#Traces	Acc.
dws	h0_v0	49	0.7589	68	0.66151	105	0.75933	48	0.68121
gt	h0_v0	37	0.91353	53	0.71118	69	0.7945	33	0.82601
	h100_v50	39	0.92279	46	0.76928	55	0.85404	26	0.84418

Table 6.2: Accuracy (average number of correct detection windows for each trace) of the modelling phase.

6.3.3 Modelling and Profiling

We evaluated the capability of the proposed global modelling solution of creating traces (e.g. a composition of tracklets) containing only the detections of the same person. In particular, we considered the detection windows obtained with our person detector (dws) or the detection windows annotated in the ground truth (gt). In this latter case, we performed a set of experiments with the full set of detection windows (h0_v0) or considering only the *reasonable* set (h100_v50). Results are reported in Table 6.2. Better performance are achieved with gt data due to the perfect alignment of the persons in the windows, with a very reduced portion of the background in the images. In this case, the re-identification of the persons is more accurate, resulting in a better creation of the traces. However, we obtained comparable results considering the h0_v0 set. This proves the effectiveness of our method to cluster together the detections of the same person.

6.3.4 User profiling

We performed a set of experiments to assess the accuracy of user profiling of Mnemosyne. For the test we considered 10 artworks of interest inside the Donatello's Hall. For each visitor, we have annotated frame by frame the observed artwork (or if no relevant object is observed by the person). We measure the accuracy of correct visitor-artwork association. In general, the performance improves using the distance and gaze together in the eq. 6.3, reaching the highest accuracy of 87% with $\alpha = 0.75$.

6.4 Conclusion

In this chapter we have presented Mnemosyne as a new system for personalized multimedia museum experience. Mnemosyne estimates the user's

interest profile based on the passive observation of visitors in museum exhibits. Visual profiling is obtained by running a complex execution pipeline composed of different computer vision task. For each video stream the system performs detection and description of the visitors in order to localize all persons in the frames, capturing the visual characteristics of each one. Then, the local modelling phase is performed to build the profiles of the visitors, separately for each camera. Finally, the global modelling merges all profiles between the cameras. The proposed strategies described in this work allow to run the pipeline in real-time.

Chapter 7

Conclusion

This chapter summarizes the contribution of the thesis and discusses avenues for future research.

7.1 Summary of contribution

This thesis makes different contributions to person detection, coarse gaze estimation and user interest profiling. We have proposed two methods to reduce the complexity of a multi-scale person detection, which address the two fundamental bottlenecks of cascade detectors: the number of weak classifiers that need to be evaluated in each cascade, and the total number of detection windows that must be evaluated. As regards the task of people profiling, we proposed a strategy to profile the attention of people moving in a known space, exploiting coarse gaze estimation and a novel model based on optical flow to improve attention prediction, without the need of a tracker.

The major contributions are summarized below:

- In Chapter 3, we proposed two strategies to reduce the computational complexity of a multi-scale pedestrian detector. Both strategies are unsupervised, based only on region of support information measured on a training set of unlabeled images. Our first strategy linearly approximates soft cascades so that only a fraction of stages must be evaluated in order to obtain an output of the entire cascade. The second strategy instead builds a generative model for candidate window proposal in order to reduce the number of infeasible windows evaluated. The ex-

periments demonstrate that both techniques are effective at increasing the efficiency of detection while sacrificing little in terms of accuracy. Linear cascade approximation yields modest improvement in efficiency. Candidate window proposal instead yields significant gains since it reduces the total number of candidate detection windows considered.

- In Chapter 4, we have proposed Mixture-based Windows Proposal (MWP) as a new method that exploits a previously learnt scene model to provide sparse proposals of windows where targets should be detected most likely. The model of the scene is built using a Mixture of Gaussians whose components take into account position, scale and relative importance of detections in their Region of Support. Experimental results have shown that detection with MWP scores the lowest miss rate with respect to state of the art methods for sparse windows proposal. This is mainly due to the ability of the method to select windows at the right scale of the target, so reducing the number of false positives.

- In Chapter 5, we have presented a solution for coarse gaze estimation that can be exploited to understand where people attention focuses. We proposed to fuse head and body orientations with a novel model based on optical flow in order to improve attention prediction without the need of a tracker. The proposed method obtains comparable performance with respect to state of the art solutions. We also show that our approach improves both mean absolute angular error and profiling accuracy on the more challenging MuseumVisitors dataset, confirming that a good coarse gaze estimate is a valuable cue for user interest profiling.

- In Chapter 6, we presented the Mnemosyne system, which makes use of passive observation to estimate the visitor's preferences in the Donatello's Hall, at the Nationl Museum of Bargello. Based on passive observation of the tourists during the visit, the system extracts a profile of interests. This is obtained by analyzing, in real-time, the video streams of four fixed cameras opportunely positioned in the hall. At the end of the visit, the tourist can interact with a user interface that shows personalized contents according to the estimated global profile. The summary of his visit can be downloaded through a mobile application, with additional multimedia contents.

7.2 Directions for future work

Recents acts of terrorism and violence have brought to light the issue of security in public areas, stations and airports. Automatic people profiling can be exploited for this purpose. In this context, people detection and re-identification have received a lot of attention due to the possibility to identify an individual over different non-overlapping cameras and viewpoints, finding the best matching in a large database of billions of people in the world In the next years I will investigate possible solutions to pedestrian detection, tracking and recognition in the wild, aiming at the development of a system that supports advanced surveillance systems with identification of the subject(s) in the frames. State of the art Convolutional Neural Network technology will be considered as subject of investigation. Since the effectiveness of this technology is strictly related to the availability of a huge number of training samples, I will also investigate the possibility of using tracking as an unsupervised method to collect different appearances of the individual's silhouette. Finally, I will analyze solutions based on fixed and first-person vision, considering technological requirements for real-time performance

Appendix A

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

In this appendix 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 1280×800 *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.*

A.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 [41] datasets are a reference for the community. Recently the large scale taxonomy annotated dataset ImageNet [32] provided the sufficient amount of data to train large and deep neural networks [56]. 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 [23, 45].

Large scale action recognition with trimmed and untrimmed videos have been recently proposed [81] 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 [5, 27]. 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 [11, 74]. More recently researchers began to address the analysis of collective patterns. A typical task is the detection of F-formations [29]; 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 [17, 24].

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 computer 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 work 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 (WATSS) 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.

A.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.

Dataset	# cam.	# frames	# ind.	# ped.	density	Group	Person ID	Occl.	Gaze	Video	Calib.
MuseumVisitors	3	4808	43	53389	11.1	✓	✓	✓	✓	✓	✓
CAVIAR Shop. Center [22]	2	72515	∼237[8]	179283	2.5	✓	✓			✓	✓
Friends meet [11]	1	10685	–	–	–	✓	✓			✓	✓
Caltech [36]	1	250000	2300	∼ 350000	1.4	✓	✓	✓		✓	
Daimler Ped. Det. [39]		21790		88880	4.1			✓		✓	
CVC-05 Part. Occl. [67]		593		2008	3.4			✓			
CUHK occlusion [72]		1063		10191	9.6			✓		✓	

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

Group detection The CAVIAR dataset [22][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 [7, 95]. 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 [11] 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 A.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 [30, 36].

The Images of Groups Dataset [44][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 [30, 36] 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.

[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

Occlusion detection Recently a lot of techniques have been focusing on person detection with occlusions handling [68, 71, 90]. 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 [39][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 [36][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, most of the pedestrians (70%) are occluded in at least one frame.

CVC-05 Partially Occluded Pedestrian dataset [67][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 [72][7] for activity and crowded scenes analysis contains 1063 images divided in 10 clips with occluded pedestrians from other five datasets: Caltech [36], ETHZ [40], TUD-Brussels [91], INRIA [30], CAVIAR [22]. 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 A.1. Here, for each dataset, we report some quantitative information: the number of cameras used (# cameras), the number of frames (# frames),

[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/

[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.

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).

A.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.

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

A.3.1 Dataset acquisition

The installation at the Bargello Museum, depicted in Figure A.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 [53]. 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.

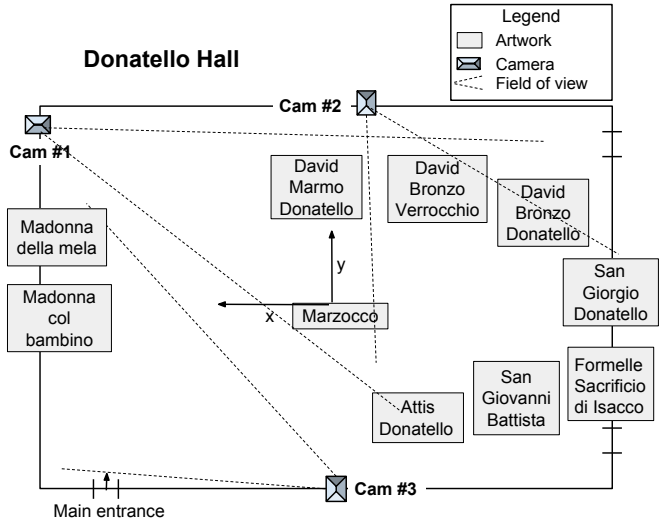


Figure A.1: Scheme of the installation at the Bargello Museum with the 3 cameras positions and fields of view, artworks location and common ground plane axis.

A.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 A.2(a). Each person is 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 A.2(b).

A.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: individ-

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

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

ual and groups. The first scenario shows visitors watching different artworks, while the second one shows groups of visitors watching the same artworks. Figure A.3 shows some sample frames for the different cameras and scenarios of the MuseumVisitors dataset. In Table A.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 [14, 30, 34, 35, 37, 42, 92]. One of the first successful approach to object detection has been proposed by Dalal *et al.* [30], 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.* [42] 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 [14, 34, 35, 37]. This recent line of work obtain state of the art performances on challenging datasets [36] and lean towards efficiency. In [35] 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 [34] (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 [14] 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 [34] moves the feature extraction complexity from test time to training time. In [37] the authors proposed the Aggregate Channel Feature (ACF) extending the work in [34] with a variant of integral

channel features to compute the pyramid features efficiently. The ACF detector was recently extended in [92] 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 [30] (I), on the Caltech pedestrian dataset [36] (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 A.4(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 A.4(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 A.4(b)) and this can be noticed from the fact that the performance of tested detectors decreases according to the occlusion level percentage, see Figure A.4(d). With this result in mind and also inspired by [36] 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 Tables A.3,A.4,A.5 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

Detector	Camera 1			
	Ind.	Ind. Reas.	Groups	Groups Reas.
HOG (I)	91	88	99	96
DPM (I)	75	69	**89**	77
ChnFtrs (I)	**67**	**57**	90	74
FPDW (I)	**67**	58	**89**	**72**
VeryFast (I+C)	95	94	98	94
ACF (I)	75	70	91	80
ACF (C)	98	93	100	96
ACF-LDCF (I)	72	65	**89**	75
ACF-LDCF (C)	93	91	98	96

Table A.3: Miss Rates @10^{-1} False Positive per Image (fppi) of leading pedestrian detectors on the MuseumVisitors dataset, for the camera 1. 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.

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 A.5 separately for individuals and groups and for each camera considering the Full scenario. While in Figure A.6 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.5 Conclusion

In this work 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

Detector	Camera 2			
	Ind.	Ind. Reas.	Groups	Groups Reas.
HOG (I)	89	80	98	97
DPM (I)	58	37	52	41
ChnFtrs (I)	**51**	**29**	42	32
FPDW (I)	**51**	**29**	**41**	**31**
VeryFast (I+C)	82	72	88	82
ACF (I)	58	48	55	47
ACF (C)	85	79	90	88
ACF-LDCF (I)	**51**	36	47	38
ACF-LDCF (C)	82	74	75	70

Table A.4: Miss Rates @10^{-1} False Positive per Image (fppi) of leading pedestrian detectors on the MuseumVisitors dataset, for the camera 2.

Detector	Camera 3			
	Ind.	Ind. Reas.	Groups	Groups Reas.
HOG (I)	95	93	100	99
DPM (I)	**32**	24	**60**	**45**
ChnFtrs (I)	37	27	73	60
FPDW (I)	51	42	75	62
VeryFast (I+C)	80	76	88	82
ACF (I)	44	38	73	62
ACF (C)	84	82	91	86
ACF-LDCF (I)	34	**23**	64	49
ACF-LDCF (C)	75	70	90	85

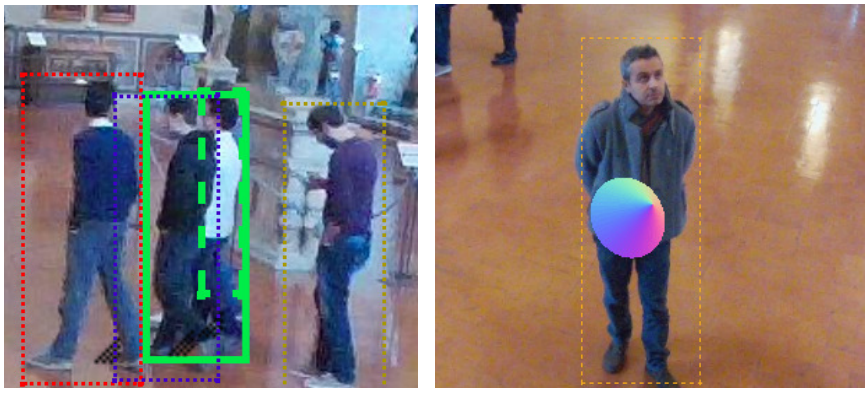
Table A.5: Miss Rates @10^{-1} False Positive per Image (fppi) of leading pedestrian detectors on the MuseumVisitors dataset, for the camera 3.

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.

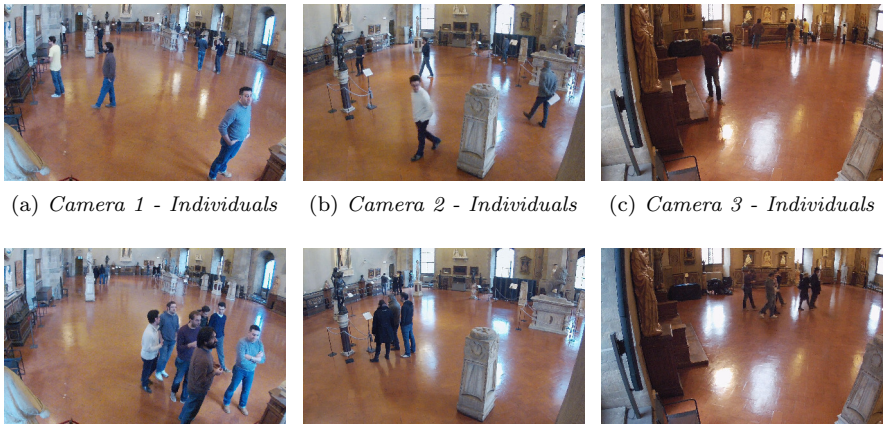
The dataset footage has been captured from a real system installed in a major Museum of the city of Florence providing challenging crowding and lighting conditions. This setup will allow us to gather more sequences in the future and release subsequent, enlarged, versions of the MuseumVisitors dataset.

Having developed a user friendly, multi-user, web based annotation tool, namely WATSS (we will present it in the next appendix), we are able to do a continuous annotation of the footage we have acquired and we have yet to release.



(a) (b)

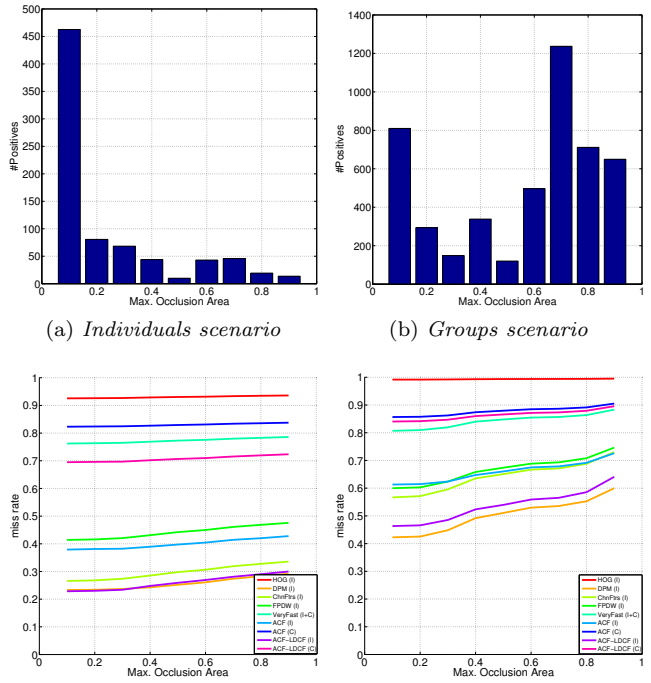
Figure A.2: *(a)* The solid green rectangle represent the bounding box selected for the annotation while the green dashed rectangle represent the visible (not occluded) area annotated by the user; *(b)* The cone visualizes the annotation of the gaze provided by the user.



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

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

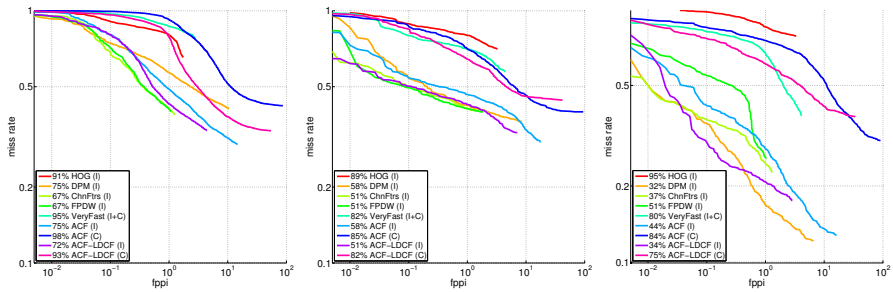
Figure A.3: Sample frames showing the different cameras and scenarios of the MuseumVisitors dataset.



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

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

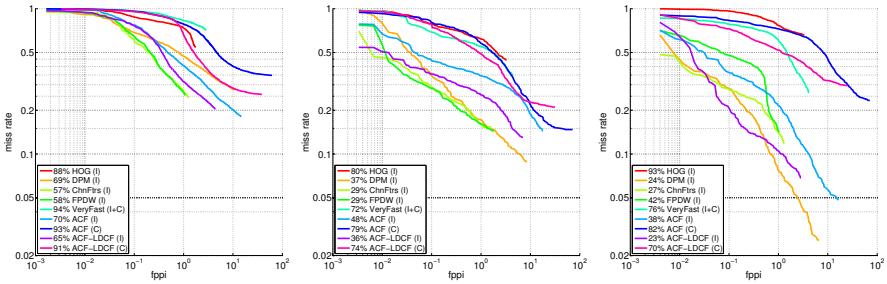
Figure A.4: Number of bounding boxes for both the individuals (a) and groups (b) scenarios for all the cameras varying the occlusion area. Average miss rate @10^{-1} averaged over the three cameras for both individuals (c) and groups (d).



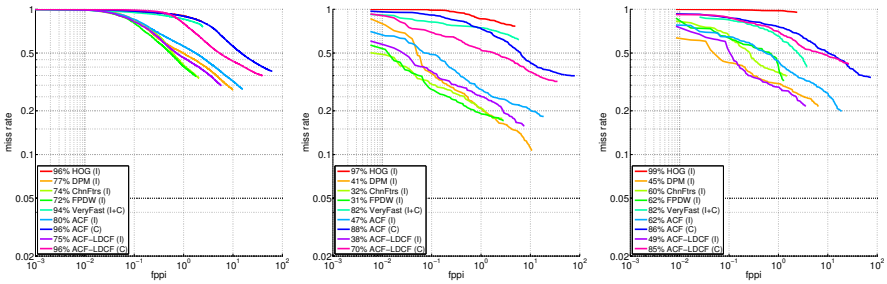
(a) *Camera 1 - Individuals* (b) *Camera 2 - Individuals* (c) *Camera 3 - Individuals*

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

Figure A.5: Evaluation results for the three cameras, on individuals and groups scenarios over all the dataset.



(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 A.6: Evaluation results for the three cameras, on individuals and groups scenarios only over the reasonable annotations.

Appendix B

WATSS: a Web Annotation Tool for Surveillance Scenarios

In this appendix, we present a web based annotation tool we developed allowing creating collaboratively a detailed ground truth for datasets related to visual surveillance and behavior understanding. The system persistence is based on a relational database and the user interface is designed using HTML5, Javascript and CSS. Our tool can easily manage datasets with multiple cameras. It allows annotating a person location in the image, its identity, its body and head gaze, as well as a potential occlusion or group membership. We justify each annotation type with regards to current trends of research in the computer vision community. We further detail how our interface can be used to annotate each of these annotations type. We conclude with an usability evaluation of our system.

B.1 Introduction

The computer vision and pattern recognition community is always seeking more challenging and realistic datasets to work on. Such datasets have been the main driver of recent major advancements in machine learning and pattern recognition. Challenges, associated with public datasets have also pushed researchers to develop methods to go beyond the state-of-the-art. PASCAL VOC [41] had been and is still advancing the accuracy of object recognition, detection and segmentation. A major break-through in image recognition has been recently made possible thanks to the large ImageNet

taxonomy [32] allowing to train a deep convolutional neural network with a sufficient amount of data [56].

Recently, researchers started to address the problem of group behavior understanding. Collective behavior understanding, like standing in groups or queuing up has been addressed in [5, 27]. The problem of person to person interaction has been tackled in [11, 74] showing that modelling social behavior can improve tracking performance. Dataset to study group behavior will often be recorded in mildly crowded environments therefore knowing whether a body is fully visible or partially occluded allows to evaluate how the methods are able to cope with occlusion. Moreover, one of the most important social cue is gaze, usually defined as a coarse gaze by the head pose since it is often not possible to detect the real gaze of a person from far field camera.

Dataset annotation is a time consuming and expensive task to perform. Recently large datasets have been annotated with crowd sourcing. Crowd sourcing usually relies on platforms like Amazon Mechanical Turk (AMT), where "turkers" are paid to perform annotations. To properly exploit AMT web based annotation interfaces [79, 88] are needed.

In this work we present an open-source tool we have developed to annotate the MuseumVisitors dataset [10]. This dataset of person and group behavior understanding, can be used for tracking, detection and coarse gaze estimation. We recorded this dataset at the National Museum of Bargello in Florence, Italy as part of the MNEMOSYNE project [53]. We designed the tool as a web application in order to easily gather annotations from multiple users and to allow concurrent annotations. The tool had to deal with multiple kinds of information thus needing a user interface designed specifically for the task. Annotators can insert groups and people identities, gaze and body occlusion.

B.2 Related tools and datasets

In this section we first review some publicly released annotation tools and then discuss the related datasets limitations that triggered the development of the WATSS tool.

B.2.1 Annotation tools

The LabelMe annotation tool [79] is focused on annotating scenes providing web based tools and mobile applications to annotate, using polygons, the outline of objects. Tools to annotate surveillance videos have been recently proposed such as VIPER [66] and VATIC [88]. These tools usually support annotations like bounding boxes, polygons and ellipses, as they are mostly developed for object detection. VATIC allows to specify a finite set of attribute per every object such as "walking" for "person" objects. The main drawback of a tool like VIPER is that is meant to be used locally instead of online, therefore the gathering of annotations from multiple sources can become difficult and there is no way of connecting the tool with crowdsourcing platforms. VATIC is a more modern online tool that can be used for crowdsourcing at scale, although their data model is extremely focused on detection and structured detection of objects [96]. The possibility to add attributes gives some flexibility to the data model but is not enough to manage the diversity of data needed for behavior understanding.

B.2.2 Group and occlusion detection datasets

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. Moreover, very few datasets have gaze annotation. In this section we briefly review some currently available datasets that contain groups or occlusion annotations.

Group detection The CAVIAR dataset [22] 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 from different cameras. These sequences were recorded acting out different scenarios of interest for different behaviors. It comes with groups annotations and it can be exploited for group detection, tracking or behavior analysis.

The Friends Meet (FM) dataset was recently proposed in [11] 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 of 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.

Occlusion detection Recently a lot of techniques have been focusing on person detection with occlusions handling [68, 71, 90]. However, due to the lack of datasets with occlusion annotations it is difficult to produce a quantitative measure of this phenomenon and compare with other methods. The Daimler Pedestrian Detection Benchmark dataset [39] 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 [36] is composed of 250000 frames extracted from 10 hours of videos acquired from a vehicle driving through regular traffic in an urban environment.

B.3 WATSS Annotation Tool

Most of current datasets are targeted for a single task, such as: person detection with occlusion, group detection and/or behavior analysis. Moreover, to the best of our knowledge no open source annotation tools are available to easily produce all the annotations needed to build a dataset covering jointly all these tasks.

We hence developed a web-based annotation tool to annotate our MuseumVisitors Dataset [10] and we made the source code publicly available. This dataset is a great example of what is needed in a modern visual surveillance dataset. In our case we want as much information as possible so we developed functionalities to annotate position, person identity, gaze, occlusion persons and group membership.

B.3.1 Annotation protocol

We propose the following annotation protocol. First of all people bounding boxes must be defined, a bounding box can be positioned and rescaled to better fit a person. If a person is partially occluded, a secondary bounding box annotation corresponding only to the visible part of the person can be defined.

Annotators can provide identities for pedestrians associating a single identifier on all frames of all cameras. Identities are easily assigned thanks

to our *Add person* interface showing avatars of already enrolled identities as show in Figure B.2.

In presence of groups, annotators can also associate a group identifier that is common to all frames of all cameras. Finally, it is possible to specify body orientation and gaze with a quantization of 5 degrees.

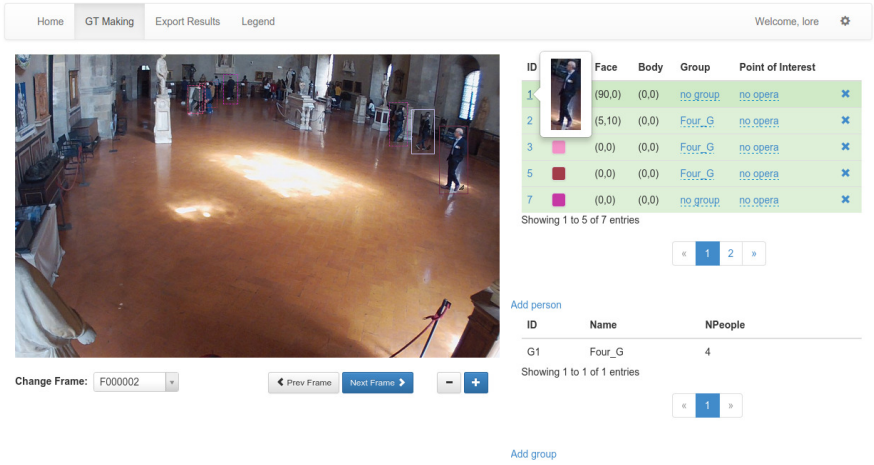


Figure B.1: Web interface. Showing several persons bounding boxes and the pop-up avatar for the first one.

B.3.2 The web based annotation tool

We designed a user friendly web interface to ease the tedious task of a detailed surveillance videos annotation. Implementing the tool as a web platform allows concurrent annotation. In fact, multiple annotators can be easily tasked with a different range of frames to annotate. Moreover the interface implement a function to point an annotator to the next un-annotated frame. In Figure B.1 we show the interface.

On the top of the interface we have a menu bar with different options: *GTmaking*, *Export results* and *Legend*. If a user selects *GT making* the annotation tool asks for 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.

On the *left-top* part of the interface, we show the chosen frame along with some already annotated bounding boxes. By selecting one of the bounding

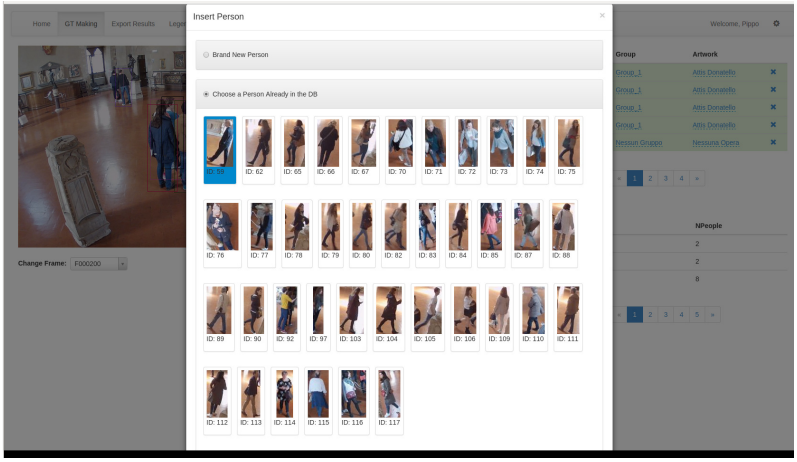


Figure B.2: Add person view. Annotators can add a new identity or select from a one previously inserted.

boxes the dashed rectangle become solid and the user is able to move and resize the bounding box. Once a bounding box is selected the user can also specify different information about that annotation, such as: the visible area (occlusion), the direction of the body and the gaze. A new bounding boxes can be added by clicking "Add person".

On the *left-bottom* part of the interface, we put some video related buttons that allows to navigate through the frames and zoom-in or out on the image (annotators can zoom also by scrolling with the mouse or touchpad).

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 object in the scene or not (Object).

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). A new group can be added by clicking "Add group" below the table.

The tool now supports CSV exporting, clicking on export data triggers the generation of an archive containing the CSV files with the annotated

data.

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 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.

B.3.3 Usability evaluation

To evaluate the usability of the proposed annotation tool we used the System Usability Scale (SUS) [20], which is a Likert scale. The form to create a Likert [84] scale is built by presenting a set of questions and asking the respondent to choose a degree of agreement in a fixed point scale, from strongly disagree to strongly agree (in our case 1 to 5). It is not just a forced choice questionnaire. Questions are selected in order to present extreme cases and alternating positive and negative statements. The alternation of positive and negative statements is a way of making sure that the respondent reads carefully. The selection of extreme scenarios is instead a way of removing bias. The SUS questionnaire was build selecting among a pool of 50 questions, those leading to the most extreme responses.

We report our usability study result in Table B.1. As suggested by Nielsen [69] five system users are enough to find the 85% of usability issues of interfaces. Regarding the SUS score our system obtained an average score of 70. We noted that all users found to be confident using the system (item 10), and the system easy to learn and use (items 2,3). We also found that many user gave a neutral response to item 6; this is probably caused by the diversity of annotations requested, but it is also room for improvement.

B.4 Conclusion

We presented a web annotation system designed for annotating multi-camera video sequences typical of surveillance scenarios.

We tested WATSS annotating our publicly released MuseumVisitors dataset comprised of 96972 detections, and gazes, 101 persons' identities over 9477 frames from four cameras. This is the work of 5 people performed through our interface for 20 days: roughly 3 man/months. We evaluated the

	Str. Dis.	Dis.	Neutr.	Agr.	Str. Agr.
1. I think that I would like to use this system to perform an annotation task	0	0	2	2	1
2. I imagine that most people would learn to use this system very quickly	1	**3**	0	1	0
3. I found the system very cumbersome to use	2	2	0	1	0
4. I thought the system was easy to use	0	1	0	**4**	0
5. I think that I would need the help of a technical person to use this system	2	2	1	0	0
6. I found the various functions in this system were well integrated	0	0	**4**	1	0
7. I thought there was too much inconsistency in this system	2	2	0	1	0
8. I found the system unnecessarily complex	2	2	0	1	0
9. I needed to learn a lot of things before I could get going with this system	2	2	0	0	1
10. I felt very confident using the system	1	0	1	2	1

Table B.1: Result of our SUS usability study. We report frequencies of each answers. Most frequent items are reported in bold.

system usability using the well known SUS scale finding that the system is considered easy to learn and use and annotators felt productive and confident in using it.

The tool is available on bitbucket at `https://bitbucket.org/fbert/watss`[1] under GPLv3 License. We provide installation scripts to feed frames into the system that can be tested at `http://150.217.35.152/watss`. We release our MuseumVisitors dataset together with the tool so that annotations can be visualized on a real world scenario.

With respect to a tool like VATIC we have a specific interface to annotate occlusions and user gaze. Moreover we are able to easily annotate user identity by showing the annotator previous persons frames. Our system provides suggestions for bounding boxes and gazes for subsequent frames so that annotators have to perform a simpler tuning task instead of redefining all scene entities from scratch. Considering the complexity of the scenarios usually involved we are not able, at the moment, to allow the interpolation of coarsely annotated sequences via tracking as in [88]. We plan in the future to add more sensible proposals for un-annotated frames both for gaze and detections in order to reduce the complexity of the annotation process.

[1] Direct download: `https://goo.gl/cgihhr`

Appendix C

Publications

This research activity has led to several publications in international journals and conferences. These are summarized below.[1]

International Journals

Submitted

1. **F. Bartoli**, G. Lisanti, S. Karaman, A. Del Bimbo. "Scene-dependent Proposals for Efficient Person Detection", *Pattern Recognition*, 2016.

International Conferences and Workshops

1. **F. Bartoli**, G. Lisanti, S. Karaman, A. D. Bagdanov, A. Del Bimbo. "Unsupervised scene adaptation for faster multi-scale pedestrian detection", in *Proc. of International Conference on Pattern Recognition (ICPR)*, Stockholm (Sweden), 2014.
2. **F. Bartoli**, G. Lisanti, L. Seidenari, S. Karaman, A. Del. Bimbo. "MuseumVisitors: A Dataset for Pedestrian and Group Detection, Gaze Estimation and Behavior Understanding", in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Boston (United States), 2015.
3. **F. Bartoli**,L. Seidenari, G. Lisanti, S. Karaman, A. Del Bimbo. "WATSS: a Web Annotation Tool for Surveillance Scenarios", in *Proc. of ACM Multime-*

[1]The author's bibliometric indices are the following: H-index = X, total number of citations = XX (source: Google Scholar on Month XX, 201x).

dia Int. Open Source Software Competition (OSSC), Brisbane (Australia), 2015.

4. I. Masi, G. Lisanti, **F. Bartoli**, A. Del Bimbo. "Person Re-identification: Theory and Best practice", in *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Washington DC (United States), 2015.

Bibliography

[1] "CAVIAR test case scenarios," http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1.

[2] "INRIA person dataset," http://pascal.inrialpes.fr/data/human/.

[3] "PETS 2009 benchmark data, dataset s2: People tracking," http://www.cvg.reading.ac.uk/PETS2009/a.html.

[4] "Pets 2009 benchmark data, dataset s2: People tracking," 2009.

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

[6] A. Angelova, A. Krizhevsky, V. Vanhoucke, A. Ogale, and D. Ferguson, "Real-time pedestrian detection with deep network cascades," in *Proc. of British Machine Vision Conference*, 2015.

[7] S.-H. Bae and K.-J. Yoon, "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning," in *Proc. of CVPR*, June 2014.

[8] F. Bartoli, G. Lisanti, S. Karaman, A. D. Bagdanov, and A. Del Bimbo, "Unsupervised scene adaptation for faster multi-scale pedestrian detection," in *Proc. of International Conference on Pattern Recognition*, 2014.

[9] F. Bartoli, G. Lisanti, L. Seidenari, S. Karaman, and A. Del Bimbo, "Museumvisitors: a dataset for pedestrian and group detection, gaze estimation and behavior understanding," in *Proc. of CVPRW*, 2015.

[10] F. Bartoli, G. Lisanti, S. Seidenari, Lorenzo Karaman, and A. Del Bimbo, "Museumvisitors: a dataset for pedestrian and group detection, gaze estimation and behavior understanding," in *Proc. of CVPR Int.'l Workshop on Group And Crowd Behavior Analysis And Understanding*, Boston, USA, 2015.

[11] L. Bazzani, V. Murino, and M. Cristani, "Decentralized particle filter for joint individual-group tracking," in *Proc. of CVPR*, 2012.

[12] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Fast stixels estimation for fast pedestrian detection," in *Proc. of Computer Vision in Vehicle Technology - ECCV Workshops*, 2012.

[13] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *ECCV*, 2014.

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

[15] B. Benfold and I. Reid, "Guiding visual surveillance by tracking human attention," in *Proc. of BMVC*, 2009.

[16] ——, "Stable multi-target tracking in real-time surveillance video," in *Proc. of Conference on Computer Vision and Pattern Recognition*, 2011.

[17] ——, "Unsupervised learning of a scene-specific coarse gaze estimator," in *Proc. of ICCV*, 2011.

[18] A. Bosch, A. Zisserman, and X. Muñoz, "Image classification using random forests and ferns," in *Proc. of ICCV*, 2007.

[19] L. Bourdev and J. Brandt, "Robust object detection via soft cascade," in *Proc. of Conference on Computer Vision and Pattern Recognition*, 2005.

[20] J. Brooke, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[21] N. J. Butko and J. R. Movellan, "Optimal scanning for faster object detection," in *In Proc. CVPR*, 2009, pp. 2751–2758.

[22] CAVIAR, "Test case scenarios," http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[23] 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.

[24] 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.

[25] C. Chen and J.-M. Odobez, "We are not contortionists: coupled adaptive learning for head and body orientation estimation in surveillance video," in *Proc. of CVPR*, 2012.

[26] X. Chen, P. Wei, W. Ke, Q. Ye, and J. Jiao, "Pedestrian detection with deep convolutional neural network," in *Computer Vision - ACCV Workshops*, 2014.

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

[28] ——, "Understanding collective activities of people from videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1242–1257, June 2014.

[29] 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.

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[31] A. Del Bimbo, G. Lisanti, I. Masi, and F. Pernici, "Person detection using temporal and geometric context with a pan tilt zoom camera," in *Proc. of ICPR*, Istanbul, Turkey, 2010.

[32] 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.

[33] P. Dollár, R. Appel, and W. Kienzle, "Crosstalk cascades for frame-rate pedestrian detection," in *Proc. of European Conference on Computer Vision*, 2012.

[34] P. Dollár, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *Proc. of British Machine Vision Conference*, 2010.

[35] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. of British Machine Vision Conference*, 2009.

[36] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *PAMI*, vol. 34, no. 4, pp. 743–761, April 2012.

[37] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[38] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, pp. 743–761, 2012.

[39] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *Proc. of CVPR*, 2010.

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

[41] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010. [Online]. Available: http://dx.doi.org/10.1007/s11263-009-0275-4

[42] P. Felzenszwalb, R. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[43] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 381–396, 2000.

[44] A. Gallagher and T. Chen, "Understanding images of groups of people," in *Proc. of CVPR*, 2009.

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

[46] P. D. Grnwald, I. J. Myung, and M. A. Pitt, *Advances in Minimum Description Length: Theory and Applications (Neural Information Processing)*. The MIT Press, 2005.

[47] G. Gualdi, A. Prati, and R. Cucchiara, "Multi-stage sampling with boosting cascades for pedestrian detection in images and videos." in *ECCV (6)*, 2010.

[48] H. Hattori, V. Naresh Boddeti, K. M. Kitani, and T. Kanade, "Learning scene-specific pedestrian detectors without real data," in *Proc. of Conference on Computer Vision and Pattern Recognition*, 2015.

[49] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in *Proc. of CVPR*, 2006.

[50] K. K. Htike and D. C. Hogg, "Efficient non-iterative domain adaptation of pedestrian detectors to video scenes," in *Proc. of International Conference on Pattern Recognition*, 2014.

[51] P. Jiang, H. Ling, J. Yu, and J. Peng, "Salient region detection by ufo: Uniqueness, focusness and objectness," in *Proc. of International Conference on Computer Vision*, 2013.

[52] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.

[53] 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*, pp. 1–25, 2014.

[54] S. Karaman, L. Giuseppe, A. D. Bagdanov, and A. Del Bimbo, "Leveraging local neighborhood topology for large scale person re-identification," *Pattern Recognition*, vol. 47, no. 12, pp. 3767 – 3778, 2014.

[55] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. of CVPR*, 2014.

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

[57] C. Lampert, M. Blaschko, and T. Hofmann, "Efficient subwindow search: A branch and bound framework for object localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2129–2142, 2009.

[58] T. Lan, Y. Wang, W. Yang, and G. Mori, "Beyond actions: Discriminative models for contextual group activities," in *Advances in Neural Information Processing Systems 23*. Curran Associates, Inc., 2010, pp. 1216–1224.

[59] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proc. of Conference on Computer Vision and Pattern Recognition*, 2014.

[60] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[61] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1629–1642, 2015.

[62] G. Lisanti, I. Masi, and A. Del Bimbo, "Matching people across camera views using kernel canonical correlation analysis," in *Proc. of ICDSC*, 2014.

[63] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. of IJCAI*, 1981.

[64] S. Manén, M. Guillaumin, and L. Van Gool, "Prime Object Proposals with Randomized Prim's Algorithm," in *Proc. of International Conference on Computer Vision*, 2013.

[65] Y. Mao and Z. Yin, "Training a scene-specific pedestrian detector using tracklets," in *Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision*, 2015.

[66] V. Mariano, J. Min, J.-H. Park, R. Kasturi, D. Mihalcik, D. Doermann, and T. Drayer, "Performance evaluation of object detection algorithms. international conference on pattern recognition," in *In Proc. of ICPR*, 2002.

[67] J. Marin, D. Vazquez, A. Lopez, J. Amores, and L. Kuncheva, "Occlusion handling via random subspace classifiers for human detection," *IEEE Transactions on Cybernetics*, vol. 44, no. 3, pp. 342–354, 2014.

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

[69] J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing*

Systems, ser. CHI '90. New York, NY, USA: ACM, 1990, pp. 249–256. [Online]. Available: http://doi.acm.org/10.1145/97243.97281

[70] N. Otsu, "A Threshold Selection Method from Gray-level Histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[71] W. Ouyang and X. Wang, "Single-pedestrian detection aided by multi-pedestrian detection," in *Proc. of CVPR*, 2013.

[72] ——, "A discriminative deep model for pedestrian detection with occlusion handling." in *Proc. of CVPR*, 2012.

[73] M. Pedersoli, J. Gonzàlez, A. D. Bagdanov, and J. J. Villanueva, "Recursive coarse-to-fine localization for fast object detection," in *Proc. of ECCV*, 2010.

[74] 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.

[75] E. Rahtu, J. Kannala, and M. B. Blaschko, "Learning a category independent object detection cascade." in *Proc. of International Conference on Computer Vision*, 2011.

[76] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proc. of CVPR*, June 2014.

[77] I. Robertson, Neiland Reid, "Estimating gaze direction from low-resolution faces in video," in *9th European Conference on Computer Vision*, 2006.

[78] N. Robertson, I. Reid, and J. Brady, "What are you looking at? gaze estimation in medium-scale images," in *Proc. of BMVCW*, 2005.

[79] B. C. Russell, A. A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation." *International Journal of Computer Vision*, vol. 77, pp. 157–173, May 2008.

[80] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1442–1468, 2014.

[81] 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.

[82] R. Sznitman, C. Becker, F. Fleuret, and P. Fua, "Fast object detection with entropy-driven evaluation," in *Proc. of Conference on Computer Vision and Pattern Recognition*, 2013.

[83] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Proc. of Conference on Computer Vision and Pattern Recognition*, 2015.

[84] W. M. Trochim *et al.*, "Likert scaling," *Research methods knowledge base*, vol. 2, 2006.

[85] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[86] P. Viola and M. Jones, "Robust real-time object detection," in *International Journal of Computer Vision*, 2001.

[87] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int. J. Comput. Vision*, vol. 63, no. 2, pp. 153–161, jul 2005.

[88] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *International Journal of Computer Vision*, pp. 1–21. [Online]. Available: http://dx.doi.org/10.1007/s11263-012-0564-1

[89] X. Wang, M. Wang, and W. Li, "Scene-specific pedestrian detection for static video surveillance," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 36, pp. 361–374, 2014.

[90] 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.

[91] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," in *Proc. of CVPR*, 2009.

[92] J. H. H. Woonhyun Nam, Piotr Dollár, "Local decorrelation for improved pedestrian detection," in *Proc. of NIPS*, 2014.

[93] Y. Xiao, C. Lu, E. Tsougenis, Y. Lu, and C.-K. Tang, "Complexity-adaptive distance metric for object proposals generation," in *Proc. of Conference on Computer Vision and Pattern Recognition*, 2015.

[94] X. Xiong and F. D. la Torre, "Supervised descent method and its applications to face alignment," in *Proc. of CVPR*, 2013.

[95] 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.

[96] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on.* IEEE, 2011, pp. 1385–1392.

[97] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *CVPR*, 2016.

[98] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *Proc. of Conference on Computer Vision and Pattern Recognition*, 2015.

[99] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. of CVPR*, 2015.

[100] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. of European Conference on Computer Vision*, 2014.