User Interest Profiling Using Tracking-free Coarse Gaze Estimation

Federico Bartoli, Giuseppe Lisanti, Lorenzo Seidenari, Alberto Del Bimbo
Media Integration and Communication Center
Università degli Studi di Firenze
Firenze, Italy
{name.lastname}@unifi.it

***Abstract*—Understanding where people attention focuses is a challenging and extremely valuable task that can be solved using computer vision technologies. In this paper 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.**

I. 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 [1]. 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 [2]. 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 [3], or perform complex and computationally onerous global optimization requiring the knowledge of the entire person trajectory.

II. Related work

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 [4], [5], [6]. An even preciser gaze estimate can be computed by locating pupils inside eyeball regions [7]. 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 [8], [9]. 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 [10].

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 [3]. 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 [11]. Their approach is shown to improve with respect to [3]. 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.

The Mnemosyne system [2] is the result of a three year research project where computer vision technologies have been deployed to perform passive user profiling. Profiles have then been used to deliver personalized content through an interactive tabletop interface or a mobile app. This system has three main components: person detection, detections association and profile building. The first step process each camera stream to locate visitors and extract their descriptors. The second step compare visitors visual descriptors to infer identities, i.e. cluster detections that are likely to belong to the same person. Profile building is performed combining a priori information about areas of influence of each artwork with visitor locations on the ground plane. Unfortunately, this approach discards completely people orientation information. In many situation a person may stand close to an artwork but look in an opposite direction.

We propose to improve profiling with respect to [2] including the estimated gaze in the profiling model. Our solution is more robust than [9] and 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. Dropping the tracker requirement is mandatory since in our scenario reliable multi-target tracking is both computationally expensive and unreliable because of the many occlusions.

III. 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. To detect person in the scene we use the detector from [12] that is able to segment both the body and the head of the detected person.

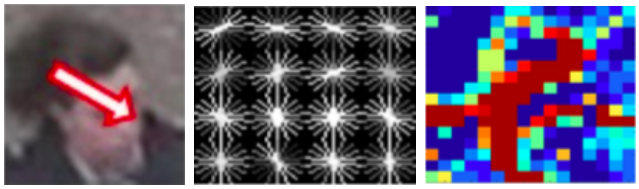


Fig. 1: Feature extracted from a sample head.

A. Head and body pose estimation

We build upon the solution proposed in [8], [9] 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. 1.

We use *random ferns* [13], as in [8], 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 [11].

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.

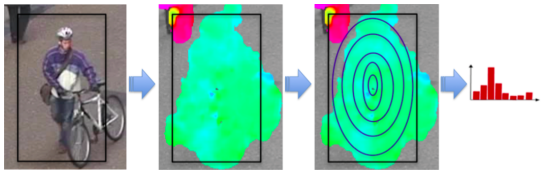


Fig. 2: Motion feature extracted from a person detection.

B. 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 [14] 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. 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 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.

IV. 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. III 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 [15] with respect to the real world reference system. To this end we first estimate

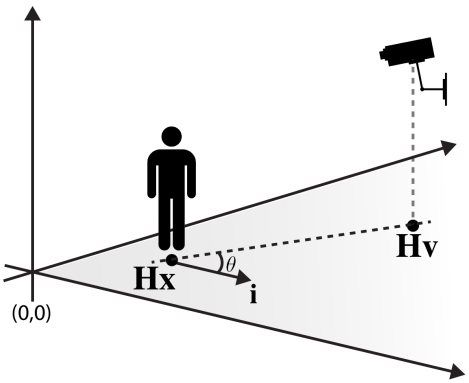


Fig. 3: Visual representation of how the compensation angle θ is computed.

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:

$$\boldsymbol{\theta} = \arccos\left(\frac{\mathbf{Hv} - \mathbf{Hx}}{||\mathbf{Hv} - \mathbf{Hx}||} \cdot \mathbf{i}\right) \tag{1}$$

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

$$d_{ik}(\alpha) = \alpha\frac{||\mathbf{p}_{ik}||}{M} + (1-\alpha)\arccos\left(\mathbf{g}_i(\boldsymbol{\theta}) \cdot \frac{\mathbf{p}_{ik}}{||\mathbf{p}_{ik}||}\right)\pi^{-1} \tag{2}$$

where

$$\mathbf{p}_{ik} = \mathbf{Hx}_k - \mathbf{Hx}_i \tag{3}$$

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:

$$k^* = \arg\min_k d_{ik}(\alpha). \tag{4}$$

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

V. 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.

A. Datasets and experimental details

Tests are conducted on two different datasets, TownCentre [9] and MuseumVisitors [16]. 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 [8], 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 [11], 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.

B. 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 I shows the performance of our strategy compared with Benfold et al. [9] and Chen et al. [11] 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. [11] 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.

Strategy	MAAE	Head Gaze	Body Gaze	Motion	Tracking
Benfold [9]	26°	✓	✗	✗	✓
Benfold [9]	26°	✓	✗	✓	✓
Chen [11]	45°	✓	✗	✗	✓
Chen [11]	28°	✓	✓	✓	✗
Chen [11]	18°	✓	✓	✓	✓
Our (O)	26°	✗	✗	✓	✗
Our (H)	42°	✓	✗	✗	✗
Our (B)	45°	✗	✓	✗	✗
Our (H+B)	42°	✓	✓	✗	✗
Our (H+B+O)	22°	✓	✓	✓	✗

TABLE I: 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	$\mathbf{26}^\circ$	$\mathbf{22}^\circ$	$\mathbf{30}^\circ$

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

In Table II 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. 4 we show 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 II, is much worse than in Table I.

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

Fig. 4: Example of persons' gaze estimated with the proposed strategy in TownCentre (a) and MuseumVisitors (b).

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

C. 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. 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 III 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 6.

Finally, in Fig 7 we show, for each camera and over all the cameras, 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.

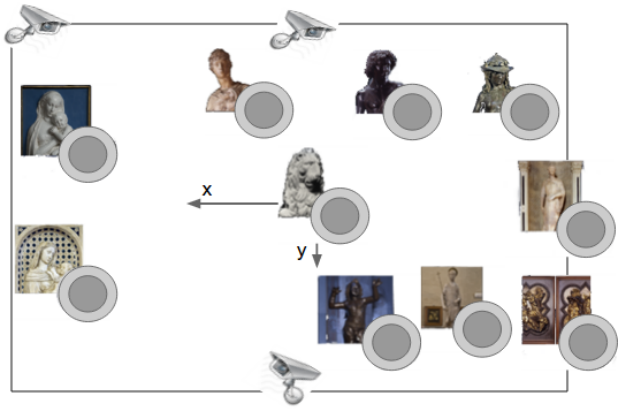


Fig. 5: Artworks location inside the Donatello's Hall.

VI. Conclusion

In this paper we 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 in terms of MAAE on the TownCentre dataset. We also show that our approach improves both MAAE and profiling accuracy on the more challenging MuseumVisitors dataset, confirming that a good coarse gaze estimate is a valuable cue for user interest profiling.

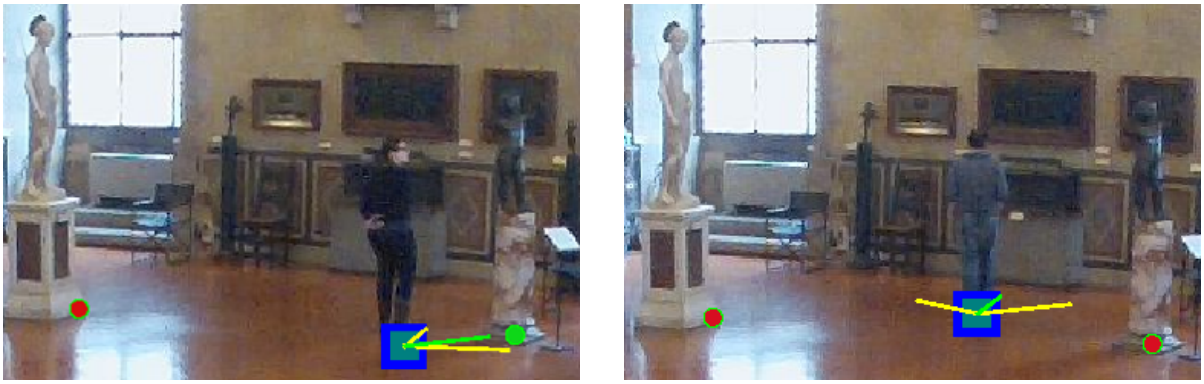
Acknowledgment

This research is partially supported by "THE SOCIAL MUSEUM AND SMART TOURISM", MIUR project no. CTN01_00034_23154_SMST.

References

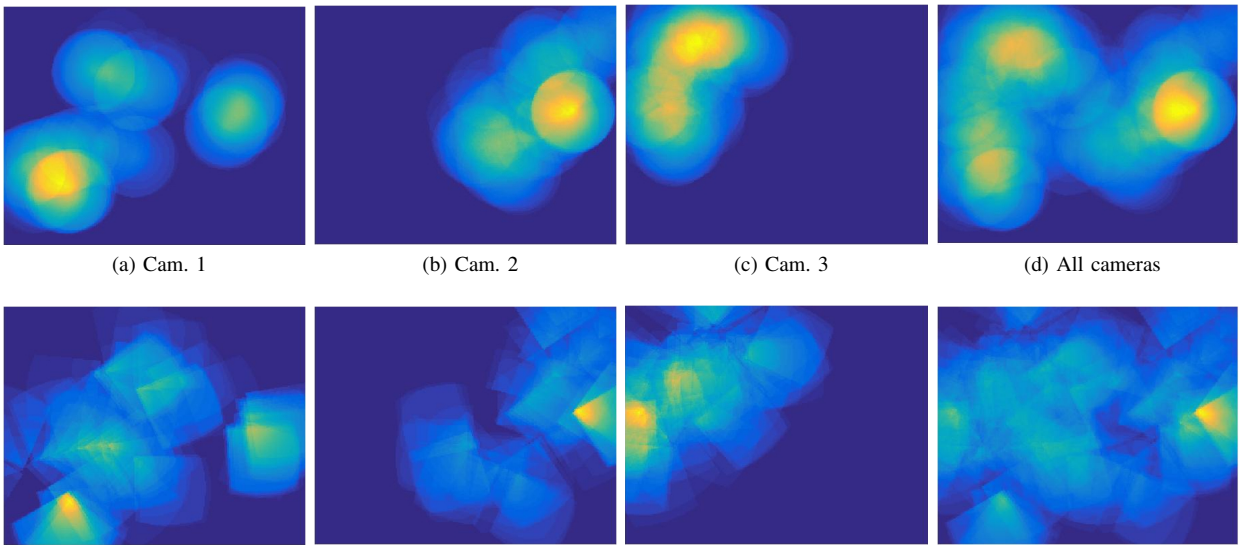
[1] G. Lisanti, I. Masi, and A. Del Bimbo, "Matching people across camera views using kernel canonical correlation analysis," in *Proc. of ICDSC*, 2014.

[2] S. Karaman, A. D. Bagdanov, L. Landucci, G. DAmico, 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.



(a) (b)

Fig. 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

Fig. 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).

[3] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in *Proc. of CVPR*, 2011.

[4] 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.

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

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

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

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

[9] B. Benfold and I. D. Reid, "Unsupervised learning of a scene-specific coarse gaze estimator." in *Proc. of ICCV*, 2011.

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

[11] 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.

[12] P. Felzenszwalb, R. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *Proc. of CVPR*, 2010.

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

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

[15] I. Robertson, Neiland Reid, in *9th European Conference on Computer Vision*, 2006.

[16] 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.