PACE: Prediction-based Annotation for Crowded Environments

Federico Bartoli, Giuseppe Lisanti, Lorenzo Seidenari, Alberto Del Bimbo
University of Florence

ABSTRACT

We present a new tool we have developed to ease the annotation of crowded environments, typical of visual surveillance datasets. Our tool is developed using HTML5 and Javascript and has two back-ends. A PHP based back-end implement the persistence using a relational database and manage the dynamic creation of pages and the authentication procedure. A python based REST server implement all the computer vision facilities to assist annotators. Our tool allows collaborative annotation of person identity, group membership, location, gaze and occluded parts. PACE supports multiple cameras and if calibration is provided the geometry is used to improve computer vision based assistance. We detail the whole interface comprising an administrative view that ease the setup of the system.

CCS CONCEPTS

• **Information systems** → *Web interfaces*; • **Computing methodologies** → *Computer vision tasks*;

GENERAL TERMS

ALGORITHMS

KEYWORDS

Annotation; Surveillance; Computer Vision

ACM Reference format:
Federico Bartoli, Giuseppe Lisanti, Lorenzo Seidenari, Alberto Del Bimbo. 2017. PACE: Prediction-based Annotation for Crowded Environments. In *Proceedings of ICMR '17, June 6–9, 2017, Bucharest, Romania, ,* 4 pages.
DOI: http://dx.doi.org/10.1145/3078971.3079020

1 INTRODUCTION

Progress in machine learning and computer vision is partially driven by the availability of large scale annotated datasets [8, 9, 12]. An important computer vision task is human behavior understanding which also addresses group behavior understanding. Not much data is available to develop learning algorithms to solve this highly challenging task. Problems such as person-to-person interaction and collective behaviors have been recently addressed [1, 5, 6, 14].

This research was supported by "THE SOCIAL MUSEUM AND SMART TOURISM", MIUR project no. CTN01_00034_23154_SMST and by "TESEO", POR FESR 2014-2020, funded by Regione Toscana, project no. 3389.30072014.068000017.
Author's addresses: F. Bartoli, G. Lisanti, L. Seidenari and A. Del Bimbo are with the Media Integration and Communication Center, Università degli Studi di Firenze, Viale Morgagni 65, 50134 Firenze, Italia. Email: name.lastname@unifi.it .

ICMR '17, June 6–9, 2017, Bucharest, Romania
 ACM ISBN 978-1-4503-4701-3/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3078971.3079020

Group behavior datasets often are collected from multiple surveillance cameras and depict crowded environments. Moreover social cues such as gaze are very relevant for interaction understanding. Since dataset annotation is extremely time consuming, collaborative distributed frame tagging is a must. Moreover supporting annotators with predictive algorithms, allowing a smooth transfer of labels from a frame to next one, is very desirable.

Thanks to the cooperation of the National Museum of Bargello of Florence and the project MNEMOSYNE we were able to collect several thousands of unannotated frames from four cameras. Part of this data has been released as the MuseumVisitors dataset [3] which we annotated using WATSS[4]. To keep up with the annotation of this evergrowing data we developed an improvement of WATSS incorporating computer vision based annotation propagation, that allows annotators to highly reduce the workload.

Our tool is web based, supports collaborative annotation and allows annotators to specify detailed information such as body and gaze orientation, identity and group membership.

2 RELATED TOOLS AND DATA

A thorough review of publicly available datasets can be found in our previous work [3, 4]. For the sake of space we do not review datasets in this paper. We now give a brief overview of existing open source annotation tools.

LabelMe [15] is a tool focused on scene annotation which also has web and mobile versions allowing collaborative work. This tool requires annotators to outline object polygons which is extremely useful for tasks such as segmentation and scene understanding but when a highly crowded environment of people has to be annotated with identity and gaze direction it is not suitable.

VIPER [13] and VATIC [16] are tools specifically devised for surveillance video annotation. VIPER is highly limited by the fact that only works as a desktop annotation thus making annotation of large dataset not scalable. VATIC is an online tool that also provide automatic prediction of annotation, but is focused on the annotation of articulated objects [17] and has some flexibility allowing attributes to be specified for every object.

In this work we build on WATSS [4], which was developed to cope with the aforementioned limitations. WATSS is a cooperative web-based tool which has similar annotation capabilities but does not provide predictive annotations and a timeline navigation. Our MuseumVisitors Dataset [3] was completely annotated using WATSS. Recently we used this data to develop a novel gaze estimation method [2]. During the annotation process we found out that to improve annotators efficiency some aid must be provided using computer vision. To the best of our knowledge VATIC is the only tool that provides such feature, however our method is geared towards multi-camera indoor visual surveillance and group behavior understanding while VATIC is a more generic tool missing some key features such as instance identity, group definition and gaze annotation.

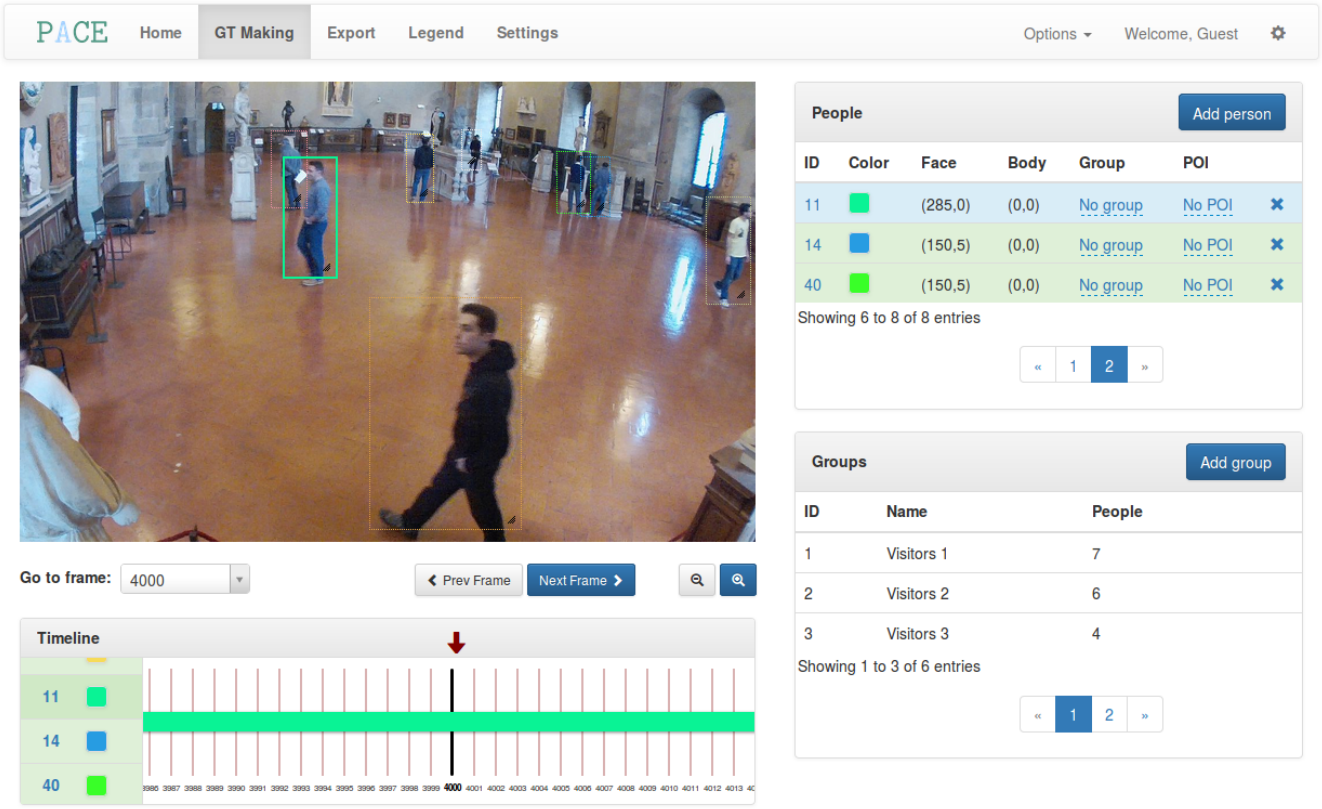

Figure 1: PACE annotation interface.

3 ANNOTATION TOOL

PACE is a web annotation tool for surveillance scenarios. The system allows annotating several information for the persons present in a frame, such as: the location (defined with a bounding box), the identity, the visible part of the body (through a second bounding box) and the orientations of both head and body. The annotation interface is shown in Figure 1.

3.1 Configuration interface

Installing a web application can be cumbersome, it often involves setting up many components such as the web server, the database and possibly other dependencies. In WATSS[4] these operation where all detailed in the documentation. At the first run, PACE will prompt the user with a setup wizard that will guide the administrative user to configure the system. This script will create the database table that will contain annotations, the configurations of every camera and users. Importing annotations previously exported from WATSS is also possible.

Moreover in PACE we added an administrative interface that checks database connection, allows to change the database password, change the camera calibration and manage users.

Figure 2: An example of frame with the timeline, showing the identity id (10) of the selected person (yellow bounding box).

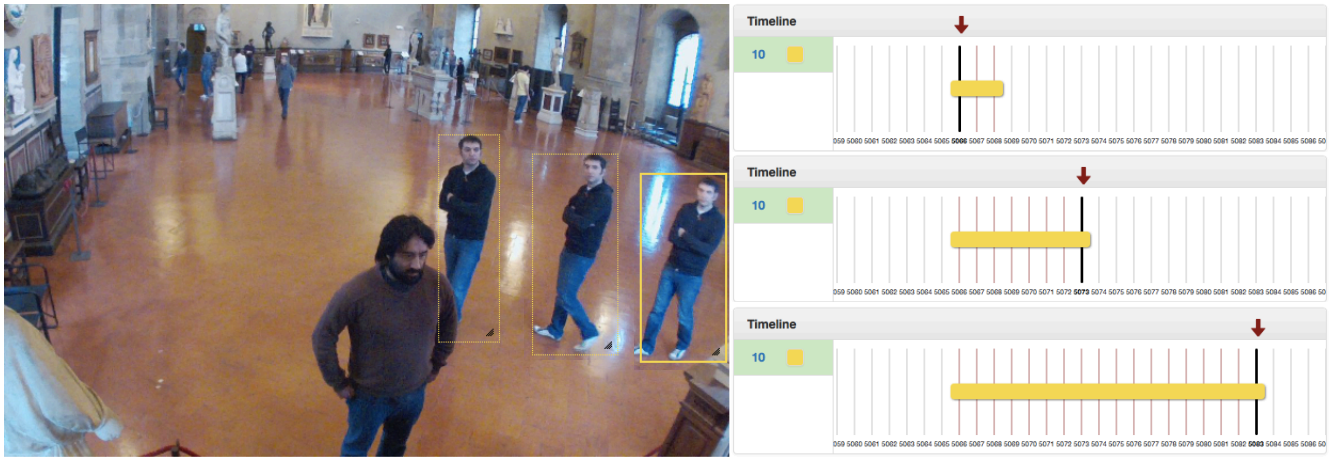

Figure 3: Example of two predictions obtained using the drag and drop function of the timeline. First, three frame are manually annotated, then the prediction is exploited to automatically generate person annotation up to 5 and then 10 frames.

3.2 Annotation protocol

PACE annotation protocol requires a user to first select to *Add Person* by choosing between a set of possible existing identities or adding a new one. The bounding box is automatically scaled depending on the position on the ground plane if the calibration is provided and the *Geometry* option is enabled (more details in section 3.5). Two bounding boxes must be specified, the whole person bounding box and the visible area bounding box. Group identities and point of interests can be also added.

3.3 The web based annotation tool

PACE interface is derived from WATSS, it is roughly split in two sections. On the right end of the screen we have two tables summarizing the persons and groups already annotated for the given frame. The top toolbar allows to navigate to the Settings and Export page. The Options menu allows to enable or disable Geometry. A link to Legend provides access to a page reporting all keyboard shortcuts.

On the left, the frame to be annotated is shown in a zoomable canvas. Zooming is extremely useful to reduce the space occupied by the frame and to allow users annotating accurately also high resolution frames. Under the frame we show the timeline which facilitates the navigation between frames and also report a list of the annotated persons for that frame. If a person is selected a bar appear over the timeline showing the annotated interval for that person, see Figure 2. The timeline interface allows the annotator to propagate the current annotation easily. Dragging the bar forward in time will trigger the computer vision based prediction, propagating annotations as shown in Figure 3.

3.4 Annotation Propagation

To ease the annotation process we extended our previous solution [4] by integrating some computer vision techniques that are used jointly to automatically generate accurate bounding box predictions for a person in the scene. In particular, given a person and its annotations from frame $t - k$ up to the current frame t we propose to infer future locations up to frame $t + m$, by exploiting motion detection, person detection and tracking.

To perform motion detection we use the background subtraction method based on Mixture of Gaussians (MoG) [10]. To this end we first train a *background model* using 40 frames randomly chosen from the whole sequence. Then for each one of the m frames we extract the foreground mask and apply morphological operations to remove small regions and merge together the one that are close to each other. Regions are then replaced with a bounding box. Bounding boxes with an area lower than a given threshold (set to 500 pixels) are discarded. A person detector based on HOG [7] is also applied to each one of the m frames in order to extract another set of candidate regions. Figure 4 shows an example of the detections process.

Given a person the system predicts the position at the next frame $t + 1$ using Kalman filtering [11]. Intersection over union scores are then computed between the Kalman prediction and the observations generated using both motion detection and pedestrian detection. The observation with the highest score is finally used to update the state of the Kalman filter and consequently the bounding box of the person. Some examples of this process are shown in Figure 5. If no valid observations are generated by both motion and pedestrian detection for that frame, the prediction of the Kalman filter is used as new bounding box.

3.5 Scene Geometry

The system also includes the possibility of specifying the geometry of the scene. In particular, through the configuration interface (see section 3.1) it is possible to specify for each camera the intrinsic parameter matrix $\mathtt{K}$, the cross-ratio parameter μ and the projective homography $\mathtt{H}$ (between the camera view and the ground plane).

Once the calibration information are specified, the scale of a person in a given location $\mathbf{z}$ can be estimated as follows:

$$\mathbf{z}' = \mathtt{W}\,\mathbf{z}. \tag{1}$$

Figure 4: Detections extraction on a sample frame *(left)* using motion detection *(middle)* and pedestrian detection *(right).*

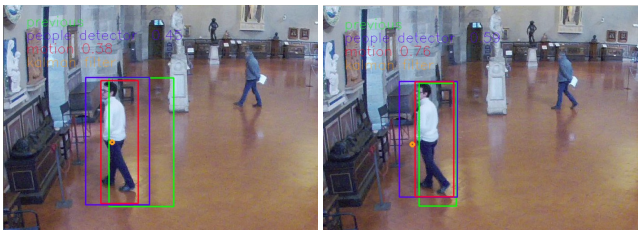

Figure 5: Example of annotation prediction for two consecutive frames. Bounding box in the previous frame *(green solid line)*, bounding box generated using motion detection *(red solid line)*, bounding box obtained through pedestrian detection *(blue solid line)*, the position predicted with the Kalman filter *(orange circle).*

being $\mathbf{z}'$ the position of the head of the person and $\mathtt{W}$ a planar homology, computed as:

$$\mathtt{W} = \mathtt{I} + (\mu - 1)\frac{\mathbf{v}_\infty \cdot \mathbf{l}_\infty^{\mathrm{T}}}{\mathbf{v}_\infty^{\mathrm{T}} \cdot \mathbf{l}_\infty}, \tag{2}$$

where $\mathbf{l}_\infty$ and $\mathbf{v}_\infty$ are respectively the vanishing line and the vanishing points, obtained as:

$$\mathbf{l}_\infty = \mathtt{H} \cdot [0, 0, 1]^{\mathrm{T}}, \qquad \mathbf{v}_\infty = \mathtt{K}\mathtt{K}^{\mathrm{T}} \cdot \mathbf{l}_\infty \tag{3}$$

When a user adds a new annotation the system exploits scene geometry to automatically suggest a dimension of the bounding box that is coherent with its perspective, making the annotation process even faster and easier.

4 CONCLUSION

We presented PACE a web application to annotate multi-camera crowded environments typical of visual surveillance datasets. The tool source can be forked or downloaded from bitbucket at https://bitbucket.org/fbert/pace and is available under GPLv3 License. Differently from WATSS, we provided an easier setup, administrative features and more importantly predictive annotations based on a computer vision engine.

REFERENCES

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

[2] Federico Bartoli, Giuseppe Lisanti, Lorenzo Seidenari, and Alberto Del Bimbo. User interest profiling using tracking-free coarse gaze estimation. 2015.

[3] Federico Bartoli, Giuseppe Lisanti, Svebor Seidenari, Lorenzo Karaman, and Alberto 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*, 2015.

[4] Federico Bartoli, Lorenzo Seidenari, Giuseppe Lisanti, Svebor Karaman, and Alberto Del Bimbo. Watts: a web annotation tool for surveillance scenarios. In *ACM Multimedia*, 2015.

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

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

[7] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proc. of CVPR*, 2005.

[8] 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.

[9] Mark Everingham, Luc Van Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, June 2010.

[10] A. B. Godbehere, A. Matsukawa, and K. Goldberg. Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation. In *2012 American Control Conference (ACC)*, pages 4305–4312, June 2012.

[11] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. of NIPS*. 2012.

[13] V.Y. 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.

[14] 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.

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

[16] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, pages 1–21.

[17] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1385–1392. IEEE, 2011.