Dana36: A Multi-Camera Image Dataset for Object Identification in Surveillance Scenarios

Janez Perš, Vildana Sulić Kenk, Rok Mandeljc, Matej Kristan, Stanislav Kovačič
Faculty of Electrical Engineering, University of Ljubljana
Trzaska 25, SI-100 Ljubljana, Slovenia
janez.pers@fe.uni-lj.si

Abstract

We present a novel dataset for evaluation of object matching and recognition methods in surveillance scenarios. Dataset consists of more than 23,000 images, depicting 15 persons and nine vehicles. A ground truth data – the identity of each person or vehicle – is provided, along with the coordinates of the bounding box in the full camera image. The dataset was acquired from 36 stationary camera views using a variety of surveillance cameras with resolutions ranging from standard VGA to three megapixel. 27 cameras observed the persons and vehicles in an outdoor environment, while the remaining nine observed the same persons indoors. The activity of persons was planned in advance; they drive the cars to the parking lot, exit the cars and walk around the building, through the main entrance, and up the stairs, towards the first floor of the building. The intended use of the dataset is performance evaluation of computer vision methods that aim to (re)identify people and objects from many different viewpoints in different environments and under variable conditions. Due to variety of camera locations, vantage points and resolutions, the dataset provides means to adjust the difficulty of the identification task in a controlled and documented manner. An interface for easy use of dataset within Matlab is provided as well, and the data is complemented by baseline results using a basic color histogram-based descriptor. While the cropped images of persons and vehicles represent the primary data in our dataset, we also provide full-frame images and a set of tracklets for each object as a courtesy to the dataset users.

1. Introduction

In computer vision, standard image and video datasets are often used to compare different algorithms in standardized and objective manner. We present a novel dataset, aimed towards computer vision methods that deal with (re)identification of objects and people in surveillance scenarios.

There are several reasons why video-based surveillance is an interesting application domain for computer vision methods. The most important reason lies in vast amounts of video data that are generated by surveillance cameras. In a large surveillance system, the number of cameras may go into thousands – for example, at the time of writing, London Underground had approximately 12,000 surveillance cameras installed. It is clear that it is impossible to manually process such amounts of data, and therefore, automatic processing methods are needed. In addition to that, video-based surveillance is regarded as a significant intrusion in the privacy of the individuals; as such, it should be correctly balanced between the degree of intrusion and the benefits to the wider society. Therefore, it is beneficial to delegate the task of "observation" to automatic algorithms, which operate without personal bias or prejudice, and are less prone to privacy subversions.

1.1. Problem statement

Surveillance can use computer-vision-based methods for many different goals. Among them are detection and tracking of people and objects, recognition of problematic or unlawful behavior, and identification of persons or objects of interest. In this paper, we focus on the problem of identification, which we define as *finding the correspondence between images of persons or objects, acquired in different instants of time, from different camera views*. In a nutshell, given a single image or small number of images of the person or object, we wish to find the same person or object across a large number of overlapping or non-overlapping camera views. This problem is distinct from pure *tracking*, which assumes more rigid temporal and, often spatial constraints, and aims to solve the problem of sequential localization and identity propagation through a*sequence* of frames.

While the problem of object *identification* is relatively

easy for human observers, it is comparatively difficult for automated algorithms, due to the following reasons:

- Variations in vantage points; resulting variations in the observed appearance are large even for the same person or object.
- Relatively small variations in appearance of different persons or objects. In surveillance scenarios we often observe small number of "natural" categories, which consist of very similarly looking instances, with many different identities. For example, a surveillance camera that supervises a parking lot, will encounter mainly people and cars, and all persons will look *roughly the same*. Combined with the variations in the observed appearance, this is in general a difficult problem, since it would ideally require a *discriminative*, not just *descriptive* model for classification. On the other hand, training a discriminative classifier on all objects that could be encountered is impossible, given the open nature of the real world applications.
- Poor quality of the real-world surveillance footage. Surveillance cameras and their placement are subject to many constraints, usually resulting in compromises regarding the resolution, viewing angle, number of cameras, lighting, etc. In the end, images are often good enough for human observers, but are of low quality compared to the data that is often used in computer-vision research.

1.2. Related work

Computer vision community has long ago recognized the need for standardized high-quality datasets. With the advent of high-speed communications, it has become an established practice for scientists to publish raw image and video data in hope that other researchers will find it useful and improve state-of-the-art results. Such datasets are usually built with a particular problem in mind, which is also reflected in their structure.

One of the basic tasks in computer vision is object recognition. The COIL-20 dataset and its more elaborate color successor, the COIL-100 dataset (containing 7,200 images of 100 objects), were developed for this purpose [1]. Both depict a number of common household items on a black background, and provide carefully controlled variations in object rotation. The datasets were aimed to demonstrate the efficiency of the parametric eigenspace technique for object recognition. Later, similar datasets, geared towards object *categorization* emerged, such as Caltech 101 [2]. It contains 101 object categories, with 40–800 images in category. The structure of Caltech 101 is much less rigid than that of the COIL-100 dataset, as it is composed of images from various sources, with no (intentional) standardization of the image acquisition. The idea behind Caltech 101 was to provide a more realistic image database that would not be biased towards any particular object recognition method. The dataset provides full images with detailed annotations (object outlines). The need for similar "unconstrained" image datasets resulted in Caltech 101 successor, Caltech 256 [3], and, more importantly, in the succession of the PASCAL Visual Object Classes challenges [4]. For those challenges, multiple image datasets for object detection and classification have been created.

Many more image datasets have been released into public use in the past 20 years. ETH80 dataset [5] is somewhat similar to COIL-100, however, it provides multiple instances of each category, foreground-background segmentation masks and includes rotation of the camera in two axes. NEC Animal Dataset [6] is also similar to COIL-100 (images of objects are acquired while they are rotated on a turntable), however, it presents a somewhat more difficult problem due to inter-class similarity (all objects are toy animals and thus have some common characteristics). Several specialized datasets emerged as well, for example Caltech-UCSD Bird 200 [7] dataset, containing 6033 images of 200 different bird species, and Flowers dataset [8], containing 102 flower categories, with 40-258 images per category. On the other hand, Middlebury dataset [9] contains stereo images and related disparity maps. However, neither Flowers nor Middlebury dataset are intended for object classification.

Surveillance datasets emerged for testing of methods developed for surveillance applications. One of more studied areas is face analysis, for which many datasets exist. For example, the CMU PIE Database [10] contains facial images in different poses, under different illumination and with different expressions. A frequent task in surveillance is vehicle detection and recognition; CMU/VASC detection test set [11] contains images of cars, and UCSD/Calit2 dataset contains images of cars with visible license plates [12]. Finally, there are several datasets that contain pedestrian images, such as Daimler pedestrian datasets [13], NICTA pedestrian dataset [14] and INRIA pedestrian dataset [15]. Those datasets are intended for pedestrian detection, which is considered especially important in driver assistance. They are not concerned with pedestrian identities.

A large amount of work on pedestrian detection, tracking and activity analysis has been done in the framework of the successive PETS workshops. However, to best of our knowledge, there are only two datasets that are specifically designed for identification and re-identification of pedestrians. First one is the VIPeR dataset [16] and the second one is the Person Reidentification dataset [17]. They provide only a small number of images from one and two cameras, respectively.

In contrast to that, when solving realistic surveillance

problems, one has to deal with large number of cameras, large number of similar objects and pedestrians, and varying image acquisition conditions. This is an especially acute problem in the field of camera networks, where researchers aim to develop solutions that would improve the performance of large, distributed camera systems, such as [18–20]. Therefore, we believe that there is a need for a large, systematically acquired surveillance dataset, which would provide all those challenges in a structured, systematic manner. In the absence of a better alternative, many of the previously-mentioned datasets can, to certain extent, be used as a substitute for surveillance footage in identification tasks. However, it has been shown in [21] that one needs a properly structured dataset with predictable and controllable level of difficulty to perform a proper evaluation in recognition tasks.

2. Dataset structure

The presented Dana36 dataset has been acquired with special attention towards providing a structured and controllable challenge for object identification algorithms. It has been devised around a synthetic scenario, which provided the same plan of activity for all participating persons and objects. In addition to the image data and annotations, it also provides Matlab API for easy integration in the research projects. It can be downloaded from the following URL (4.36 GB): `http://vision.fe.uni-lj.si/research/dana36/`

2.1. Dataset scenario

The dataset scenario is depicted in Figure 1. It models the arrival of people at a large building, and their subsequent departure. All people arrive in cars, which enter the premises at the gate, and drive to the parking lots P1 and P2. At P1 and P2, people exit the cars and walk towards the main entrance, through the lobby and up the stairs towards the first floor. After that, scenario is switched to departure: people walk in opposite direction, exit the building, enter their cars, and drive away.

2.2. Cameras and viewpoints

One of the primary goals was to create the dataset with as many camera views as possible. Unfortunately, due to equipment constraints, the recording had to be performed in multiple iterations, with a number of stationary, but partially relocatable cameras.

Resulting camera views are summarized in Table 1 and shown in Figure 2. It can be seen that there is indeed a wide variety of views. Cameras have been placed both indoors and outdoors (note that Table 1 lists the location of the observed area, not the location of the cameras – the latter is irrelevant for our study). There are three distinct

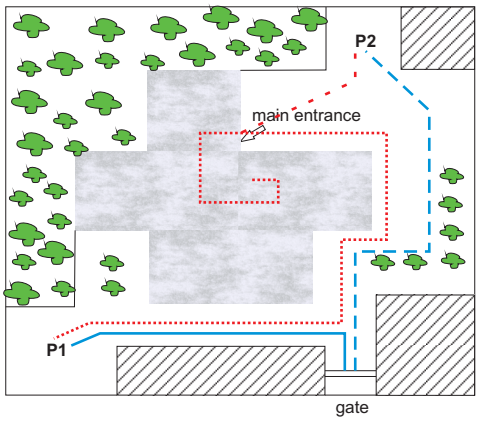

Figure 1. Dataset scenario - approximate area map and movement plan of cars and people that is depicted by the dataset. Blue lines show car paths and red lines correspond to people paths.

Table 1. Camera views

View	Location	Vantage*	Resolution	Quality*
1	Outdoor	High	1280×720	High
2–21	Outdoor	Low	640×480	Medium
22,26	Indoor	Low	640×480	Medium
23-25	Indoor	Medium	640×480	Medium
27-30	Indoor	Medium	2048×1536	High
31-33	Outdoor	High	704×576	Low**
34-36	Outdoor	High	1296×972	High

* Description of properties is provided in the text.
** Due to de-interlacing of the interlaced source video.

categories of vantage points: *low* means that cameras were slightly lower than average human height, and their view of objects of interest was basically horizontal. *Medium* vantage point means that cameras were put slightly higher than average human height, therefore they have slightly elevated view of the objects. This includes cameras that have been placed at the stairwells. Finally, *high* vantage point means that cameras have been placed significantly higher than human height, and they have significantly elevated viewpoint. Camera resolution varied from low, analog, to three megapixel. Finally, the Quality column in Table 1 denotes the *subjective* video quality, as assessed by the dataset authors. To obtain this rating, we considered quality of focus, color vibrancy and image artifacts (e.g. due to interlacing, which was not fully remedied by deinterlacing procedure). Dataset provides two Matlab functions, which can be used to select specific subset of objects and camera views in a controlled way (for example, a sub-dataset containing *only people in high-definition outdoor views*).

The dataset also provides background images for each camera view; they are shown in Figure 2. Background images may be used to perform rough segmentation of objects

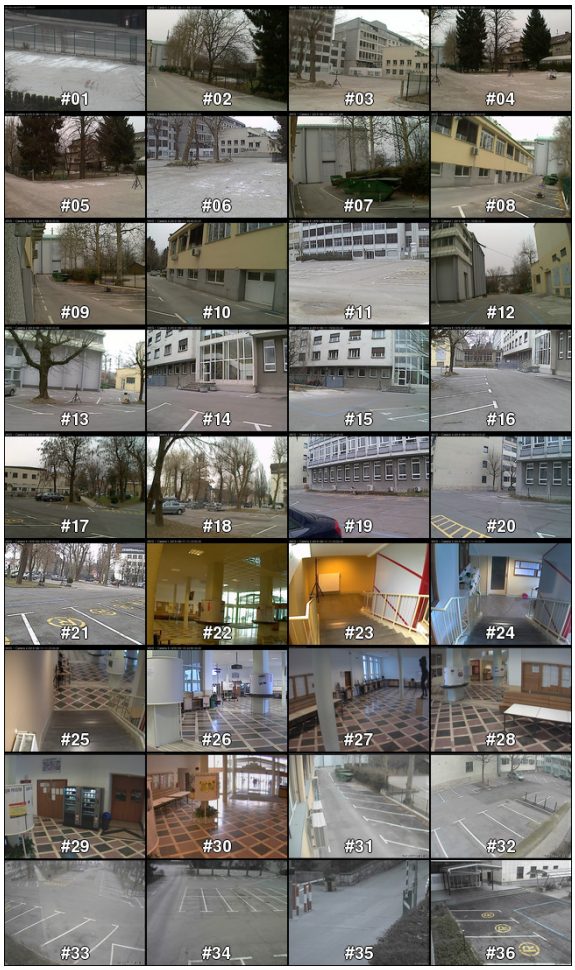

Figure 2. Background images for 36 camera views that have been used to build the dataset. Properties of the individual camera views are documented in Table 1.

from the background.

2.3. Dataset objects — people and cars

Dataset contains images of 15 people of both genders and nine cars. Recording took place during winter, and people are clothed accordingly, and they do not change clothes even when indoors. Representative images are shown in Figure 3.

2.4. Acquisition procedure

Data acquisition and preparation has been performed as follows. Fully-stationary cameras (views 1, 27–30) were set up for recording, and permission to use recordings from the existing surveillance cameras on the premises (views 31–36) was secured. Five mobile cameras (views 2–6) were set up for recording at the parking lot P1. People arrived by cars, exited, and walked past the cameras towards the entrance of the building, as shown in Figure 1, until they

Figure 3. The dataset contains images depicting 24 objects — 15 people and nine cars.

were outside the camera views. Then people returned, and drove away. Five mobile cameras were relocated to acquire views 7–11; arrival by cars, exiting and walking was repeated. The whole procedure was repeated again to generate views 12–16 and 17–21. In the last iteration, indoor scenes were recorded by a combination of both mobile (22–26) and stationary (27–30) cameras.

Where necessary, recordings were processed (e.g. deinterlacing on views 31–33) and converted to Motion JPEG format. Bounding boxes of every person or car were annotated using an annotation tool developed in Matlab, and full-frame videos were converted to a large number of JPEG files, which comprise the bulk of dataset data. Additionally, background images were generated from full-frame videos by per-pixel median operator along the temporal axis. We provide full-frame images, bounding box annotations, and on-the-fly cropping functionality, courtesy of the included Matlab dataset interface. Note that only the cropped images of people and cars are considered *primary* dataset data. Full-frame images are provided only for convenience, and the annotations on full frames may not be complete. For details, see Section 3.3. Finally, the temporal sequence of annotations from the video stream was used to automatically derive a number of short *tracklets*. Tracklets are defined as a sequence of object instances that have been annotated over an uninterrupted sequence of frames in a single view. If at any given frame the specified object has not been annotated (for any reason, e.g. either due to object disappearance or negligence on the side of the operator), the previous tracklet

ends and next tracklet begins with next annotation. We provide these as a convenience as well, and their use is subject to limitations, described in Section 3.3.

No camera calibration was performed, and no attempt to alter the temporal sequencing of images was done. Therefore, the image sequences still reflect acquisition procedure to some extent, and for that reason the dataset should not be used for any experimental work regarding camera handover algorithms or other methods that assume rigid temporal or spatial constraints.

2.5. Privacy, consent and limitations

All participants signed consent forms for video recording, however, the images from the dataset may be only used for research purposes. If any image from the dataset is published in any form (e.g. part of a research paper, presentation, or poster), *it is required* that the authors of publication blur the face and the car license plate using the function, supplied as part of the dataset's Matlab interface.

3. Intended use

The dataset is intended for use in object classification and person (re)identification tasks.

3.1. Baseline results

To establish the difficulty of the dataset and some of its peculiarities, we performed some baseline tests. We claim no innovation in this respect, and we expect that the state-of-the-art methods will be able to surpass the results presented here. We used a basic RGB color histogram descriptor with $6\times6\times6$ bins to convert each image to the corresponding 36-dimensional feature vector. Each of the 23,683 descriptors was compared to all other descriptors, yielding a diagonally-symmetric distance matrix. By varying the thresholds from zero to maximum distance, Receiver Operating Characteristic (ROC) curves were generated, and are shown in Figure 4.

By extracting the diagonally-symmetric submatrices from the distance matrix, ROC curves for the tree different subsets (only images with cars, only images with people, only indoor images) were generated as well. Although we used very basic descriptor, the results are provided to give rough difficulty estimate regarding the identification task on this dataset.

3.2. Appropriate and inappropriate uses of the dataset

Researchers are invited to use the dataset either for surveillance-related identification or general object classification problems. By providing many camera views with different characteristics and image quality, the dataset provides means to vary the difficulty of the problem in a clean

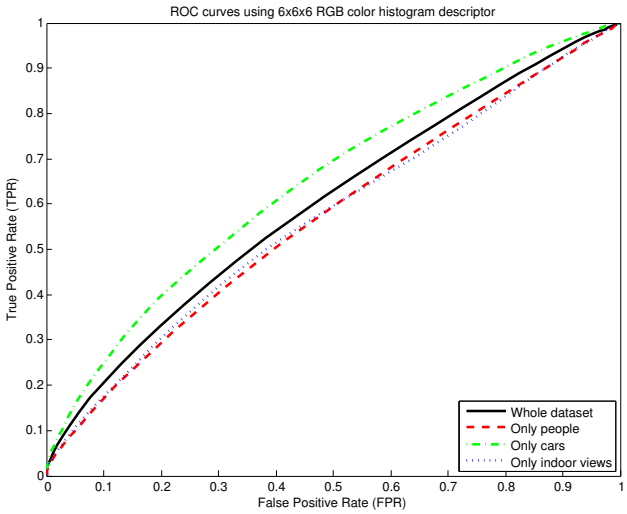

Figure 4. Performance (ROC curves) of the basic color histogram-based object-matching scheme on the whole dataset (solid, black line) and comparison with the three smaller subsets.

and documented way, so authors should report which camera views (and objects) have been used in their evaluation. We suggest that the researchers follow the methods of performance evaluation and reporting that are customary for their field of work.

It should be noted that due to the specifics of data acquisition (multiple passes), the dataset is less suitable for testing multi-camera and tracking algorithms.

3.3. Primary and secondary data

We consider cropped images, as provided by the supplied Matlab interface (which performs image cropping on the fly, to reduce redundancy in the dataset), as the *primary dataset data*. To obtain this data, great emphasis has been placed on correctness of the annotations. The cases when the identity could not be unambiguously determined or when the resulting bounding box was too small were intentionally left unannotated. Only objects with known identities have been annotated.

On the other hand, full-frame images and tracklets are considered *secondary dataset data* since they have been extracted as a side benefit of the annotation process, and are provided as a courtesy to the dataset users. Full-frame images may be used for evaluation of object detection, but with consideration: recall of the objects that are annotated in a particular frame can be measured, but precision cannot be – occasionally, an image may contain foreign objects that were not included in annotation, or, more often, the known objects may not be annotated due to oversight on the side of the operator, or due to operator's judgement that the object bounding boxes would be too small. For the same reason, tracklet information may be inaccurate, and should be man-

ually verified before using the data for serious performance analysis of object tracking algorithms.

3.4. Difficult vs. easy problems and their relevance

The interested public may test their state-of-the-art descriptors and classifiers on the presented dataset in any way they feel appropriate. Nevertheless, we present a few guidelines regarding problem difficulty to help the researchers construct a suitably challenging and relevant experiments.

In realistic surveillance scenarios there are not many images available for training, and rarely they are obtained from multiple views. The basic testing described in this paper is on the extreme end of one-shot learning, assuming that only one image of an object is known; this is the most challenging task that can be constructed using this dataset. Increasing the training set to include multiple (or even all) images from a single view should make the problem easier, and adding multiple views to the training set should decrease the difficulty even more. On the other hand, restricting the problem to lowest-quality views, as specified in Table 1, should increase both difficulty and relevance of the obtained results to the actual surveillance applications. Finally, including many images from all or most of the camera views into the training set and using k-fold cross validation (similar to the evaluation protocols for Caltech 101) should result in the least challenging problem.

4. Conclusion

We presented a novel dataset for evaluation of object matching and recognition methods in surveillance scenarios. With 36 camera views, 24 objects and 23,683 images it significantly improves the state-of-the-art in the available surveillance datasets. The variety of views and two different natural categories of objects (cars and people) enable construction of experiments with controlled and documented level of difficulty.

References

[1] S. Nayar, S. Nene, and H. Murase, "Columbia object image library (COIL 100)," Department of Comp. Science, Columbia University, Tech. Rep. CUCS-006-96, 1996.

[2] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *PAMI*, vol. 28, no. 4, pp. 594–611, 2006.

[3] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Institute of Technology, Tech. Rep. 7694, 2007.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, jun 2010.

[5] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *CVPR*, vol. 2, june 2003, pp. 409–15.

[6] H. Mobahi, R. Collobert, and J. Weston, "Deep learning from temporal coherence in video," in *ICML*, 2009, pp. 737–744.

[7] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-UCSD Birds 200," California Institute of Technology, Tech. Rep. CNS-TR-2010-001, 2010.

[8] M. Nilsback and A. Zisserman, "Delving deeper into the whorl of flower segmentation," *Image and Vision Computing*, vol. 28, no. 6, pp. 1049 – 1062, 2010.

[9] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *CVPR*, june 2007, pp. 1–8.

[10] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression database," *PAMI*, vol. 25, no. 12, pp. 1615 – 1618, December 2003.

[11] H. Schneiderman and T. Kanade, "A statistical model for 3d object detection applied to faces and cars," in *CVPR*, vol. 1, June 2000, pp. 746–751.

[12] L. Dlagnekov and S. Belongie, "UCSD/Calit2 Car License Plate, Make and Model Database," http://vision.ucsd.edu/car_data.html.

[13] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *CVPR*, june 2010, pp. 990–997.

[14] G. Overett, L. Petersson, N. Brewer, L. Andersson, and N. Pettersson, "A new pedestrian dataset for supervised learning," in *IEEE Intelligent Vehicles Symposium 2008*, june 2008, pp. 373–378.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, vol. 1, 2005, pp. 886–893.

[16] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV*, 2008, pp. 262–275.

[17] M. Hirzer, C. Beleznai, P. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *SCIA*, 2011, pp. 91–102.

[18] J. Park, P. C. Bhat, and A. C. Kak, "A look-up table based approach for solving the camera selection problem in large camera networks," in *Proc. of Int. Workshop on Distributed Smart Cameras*, 2006, pp. 72–77.

[19] A. Y. Yang, S. Maji, C. M. Christoudias, T. Darell, J. Malik, and S. S. Sastry, "Multiple-view object recognition in band-limited distributed camera networks," in *ICDSC*, 2009, pp. 1–8.

[20] V. Sulić, J. Perš, M. Kristan, and S. Kovačič, "Efficient feature distribution for object matching in visual-sensor networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 7, pp. 903–916, 2011.

[21] N. Pinto and J. J. Cox, D. amd DiCarlo, "Why is real-world visual object recognition hard?" *PLoS Computational Biology*, vol. 4, no. 1, p. e27, 01 2008.