Self-calibrating 3D context for retrieving people with luggage

Johannes Schels, Joerg Liebelt, Rainer Lienhart

Self-Calibrating 3D Context for Retrieving People with Luggage

Johannes Schels*, Joerg Liebelt*
EADS Innovation Works
München, Germany
{johannes.schels,joerg.liebelt}@eads.net

Rainer Lienhart
University of Augsburg
Augsburg, Germany
lienhart@informatik.uni-augsburg.de

Abstract

We outline the retrieval of images from a network of se-curity cameras by means of an attribute-based query. Our approach is based on detectors for several object classes which enable combined queries to retrieve people based on characteristic pieces of luggage. The approach works in-dependently of camera recording frame rates since it does not rely on tracking or background assumptions, and it re-quires neither real training images nor manual annotations since it is entirely trained on synthetic data. By performing an approximate 3D auto-calibration for each camera from a few detected humans and exploiting object-level context in a 3D coordinate system, we can significantly improve the precision of otherwise weakly performing detectors for in-conspicuous object classes. We evaluate our approach on data from an airport security camera network and demon-strate the system's ability to respond to combined appear-ance and 3D metric contextual attribute queries over multi-ple cameras.

1. Introduction

When trying to take into account the sketchy descrip-tions given by witnesses, the image-based retrieval of indi-viduals from large-scale surveillance camera records turns out to be particularly challenging. It is difficult to iden-tify people in crowded videos with PAL resolutions based on vague pieces of information such as "of medium height, wears a dark shirt". Consequently, adding every potentially available contextual piece of evidence to the query, such as descriptions of the pieces of luggage carried (e.g. "has a red backpack"), could significantly improve the query's discriminative power. However, individual detectors for relatively inconspicuous object classes still do not perform sufficiently well, as is demonstrated in the VOC2010 chal-lenge [7]. This is not surprising when looking for example at the sparse appearance and geometry clues of a backpack in figure 1.

Figure 1. Without spatial context, the two image regions are dif-ficult to interpret; this is a typical problem for detectors of incon-spicuous object classes.

Consequently, the use of object-level context has been advocated in order to exploit the co-occurrence of object classes which might otherwise not be detectable individ-ually [6]. However, their deployment in real application scenarios is frequently prevented by small object sizes and the fact that pure 2D image-space context does not pro-vide enough spatial discrimination in multi-view detection settings (see figure 2). More sophisticated 3D context re-quires information on scene geometry, but most current ap-proaches rely on restrictive prior assumptions on scene lay-out [11] or manual calibration, neither of which is accept-able for flexible surveillance tasks. In addition, low record-ing frame rates may prevent the use of tracking for the es-timation of vanishing points from ground plane trajecto-ries [14], and manually annotated training data sets con-taining all possible context configurations between object classes are usually unavailable.

In the present paper, we outline an approach which is trained on purely synthetic data without knowledge or de-pendency on a real scenario, therefore being much more

Figure 2. Spatial context for a person and a trolley: the 2D detection of the person (solid gray line) allows to derive a 2D image area in which to search for a trolley (red dashed line). However, the 2D context does not take into account 3D perspective changes (e.g. a large camera tilt as shown here). In contrast, the projected 3D context (see figure 6) models the search area more accurately (red solid line).
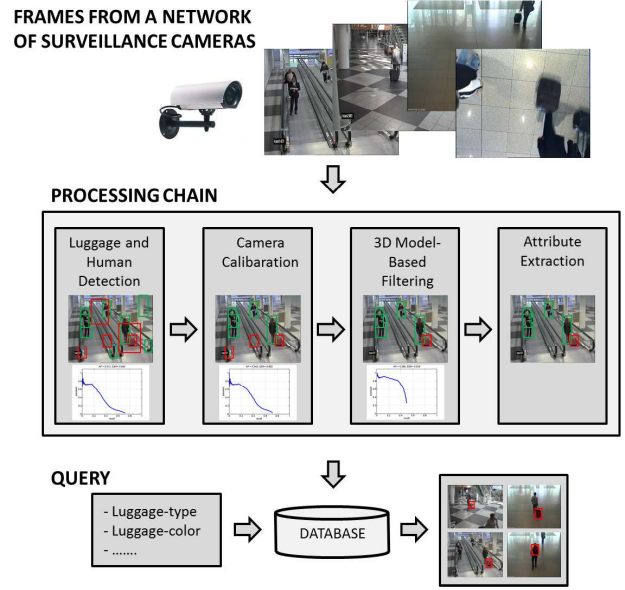

Figure 3. We apply object detectors, approximate camera auto-calibration and 3D object-level context filtering to sequences of frames from different surveillance cameras without temporal coherence. As a result, complex attribute-based retrieval queries can be answered which combine object class, appearance and metric information.

universally applicable. It does not rely on tracking and estimates the approximate 3D scene geometry for each camera from a few human detections in single frames in a fully unsupervised way. The resulting 3D geometrical context between people and different pieces of luggage such as trolleys, suitcases and backpacks, is exploited in order to improve the detection precision for these challenging object classes. Consequently, retrieval queries combining object classes, appearance clues and metric information (e.g. "tall person, wearing a dark shirt, carrying a red backpack and a trolley") can be answered efficiently. We evaluate our approach on a realistic data set from an airport surveillance camera network containing approx. 6000 images which stem from short video sequences and single frames and demonstrate a significant improvement achieved with our context approach over individual object detectors.

The paper is structured as follows: section 2 summarizes previous work on attribute-based retrieval, context-based object detection and camera calibration. An overview of the proposed approach is given in section 3. In section 4, a detailed description of the processing chain is provided. Experimental results are outlined in section 5.

2. Related Work

The present work outlines an approach to the attribute-based retrieval of individual surveillance camera frames. Numerous previous publications address similar tasks: [9] describes tracking and motion classification to perform attribute-based vehicle retrieval for calibrated cameras. [10] uses multi-layer adaboost classifiers for vehicle search which are trained on partially synthetic training data, while [20, 23] present methods where classifiers are trained for individual body parts to retrieve people via fine-grained attribute queries. Although many other approaches do not explicitly focus on retrieval, they can potentially be used to facilitate this task: e.g. [2] describes the tracking of

pedestrians and the stable association of their trajectories, whereas motion-segmented silhouettes [22] or temporal silhouette templates [5] can be used to infer if a person is carrying an object. In contrast, the present approach does not require consecutive video sequences. Instead of tracking or motion segmentation, it relies exclusively on classifiers for people and pieces of luggage which makes the approach independent of camera frame rates, background and motion assumptions. In recent years, latent part-based object detectors [8] based on HOG descriptors [4] have yielded impressive results for many object classes. In [13, 19], the use of synthetic training data has decreased the dependence on specific real training data sets and allowed systematically varying training viewpoints and imaging conditions; in the context of detecting humans, real and synthetic training results compared favourably in [15, 17]. Consequently, we build on these results in training [8] exclusively on synthetic 3D human models. Still, significantly reduced detection performances can be observed for certain inconspicuously textured and shaped object classes [7]. However, for surveillance tasks, inconspicuous object classes such as suitcases are highly relevant. In order to improve their detection results, the present paper exploits object-level context from 3D spatial co-occurrence between detectors for conspicuous classes, e.g. people, and detectors for less conspicuous but more task-relevant classes, e.g. pieces of luggage. The use of context has been advocated before to model object-to-object as well as object-to-scene dependen-

cies; an overview is given in [6]. 3D spatial context for calibrated scenes has been used in [21] to improve people detection and tracking by adapting detectors to local scene constraints; in contrast, our approach does not require prior calibration and it is therefore closer to [11] who estimate a 3D camera model to remove geometrically unlikely detections, or [1, 3, 24] who recover partial 3D geometry to incorporate the expected visibility of objects. However, [18, 11] assume prior knowledge on scene geometry, and many classical approaches to auto-calibration from object detections rely on tracking (e.g. [12]), require the computation of vanishing lines (e.g. [14]) or assume identical object heights [16]. In contrast, our approach approximates intrinsic and extrinsic camera parameters with a probabilistic approach to align the distribution of detected human heights with a known human height prior without relying on tracking. Although less precise, the approach is sufficient to provide the framework for evaluating the 3D spatial context between human and various luggage detections.

3. Overview

Figure 3 gives an overview of the system proposed in this paper. We assume a realistic scenario where frames from different uncalibrated pan-tilt-zoom surveillance cameras in a large-scale network are stored in a database for a forensic retrieval of people based on appearance, metric information and the kind of luggage they carry. To reduce network load and storage space, only single frames are recorded which prevents the use of tracking approaches relying on temporal coherence. Each frame is tagged to indicate if the operator moved the camera since the last stored frame. Our approach proceeds as follows:

1. For each camera in the network, its frames are grouped into sets without intermediate camera movement events. Object class detectors for humans and different types of luggage are applied to each frame of the set; see section 4.1. If at least one human detection is found in the current set, the next process step is triggered.

2. An approximate auto-calibration of the intrinsic and extrinsic parameters of the camera belonging to the current set is performed; see section 4.2. Note that the camera parameters are known to be constant within one set. We use a maximum likelihood estimation that aligns the height distribution of the human detections after 3D reconstruction with a known target height distribution. The estimation allows for the removal of object detections (humans and luggage) which are inconsistent with the 3D scene geometry (see figure 7, center).

3. By combining the reconstructed 3D scene geometry with a static 3D context model for humans carrying different pieces of luggage (see figure 6), we can sig-

Figure 4. Part-based object class detectors are trained on synthetic data. We only show the root filter for each class.

nificantly improve the precision of the luggage detections.

4. Meta-data from all filtering steps can now be stored in a database in order to efficiently and precisely process attribute-based queries; see the experimental section 5 and figures 10,11,12 for examples.

4. Processing Chain

In this section, we describe the processing chain of our approach.

4.1. Object Class Detector

The present work builds on the current state-of-the-art detection approach of [8] which learns discriminative multi-scale deformable part models based on HOG descriptors. In [8], a part model for an object is subdivided into a global root filter and several latent part filters which represent an object as a flexible constellation of several components. Instead of using real training data, we follow the ideas of [13, 19] and train our object class detectors with synthetic 3D object models which do not require any manual annotations and allow generating a large amount of training data. In figure 4, the resulting root model and a training example for each object class are shown. See section 5.1 for details on the data generation.

4.2. 3D Auto-Calibration

Given a set of frames from a camera without intermediate camera movement, we propose a 3D auto-calibration step in order to determine the framework for evaluating 3D object-level context. The calibration step is based on a Bayesian formulation to align the distribution of detected human heights with a known human height prior; it can be considered a generalization of the idea outlined in [11]. We describe a simplified camera model and derive a relationship between a detected person in an image and its corresponding height in the scene. We then approximate the intrinsic and extrinsic camera parameters via a maximum

likelihood estimation. The following notations will be used: world coordinates $\mathbf{X_w} = (X_w, Y_w, Z_w)$; camera coordinates $\mathbf{X_c} = (X_c, Y_c, Z_c)$; camera height Y_w^c defined in the world coordinate system; pixel coordinates $\mathbf{u} = (u, v)$; camera tilt θ; focal length f; and camera optical center (u_c, v_c).

4.2.1 Camera Model

We rely on a simplified camera model and assume zero roll and an optical center of the camera in the center of the image plane. The ground plane is defined by $Y_w = 0$ where all human detections in one set of frames per camera are assumed to have their foot points located on this ground plane. Our approach is independent of the camera yaw angle and could theoretically sustain a changing camera yaw within one set of frames, as long as the same ground plane assumption holds. We use a perspective projection model with unit aspect ratio. In homogeneous coordinates, the transformation from world coordinates to camera coordinates is given by

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & \sin\theta & -Y_w^c \cos\theta \\ 0 & -\sin\theta & \cos\theta & Y_w^c \sin\theta \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}. \quad (1)$$

The transformation from camera to pixel coordinates which defines the perspective projection Φ then has the form

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{Z_c} \begin{bmatrix} f & 0 & u_c & 0 \\ 0 & f & v_c & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix}. \quad (2)$$

From equation 1 and equation 2 we can solve for Y_w:

$$\begin{aligned} Y_w =\ & \frac{Z_w(v\cos\theta - v_c\cos\theta - f\sin\theta)}{f\cos\theta + v\sin\theta - v_c sin\theta} \\ & + \frac{Y_w^c(v\sin\theta - v_c\sin\theta + f\cos\theta)}{f\cos\theta + v\sin\theta - v_c sin\theta}. \end{aligned} \quad (3)$$

In this paper, we assume that all detected humans have foot points located on the ground plane. Given the top and bottom position of a human detection in an image, v_t and v_b, we can solve equation 3 for depth Z_w, since $Y_w = 0$ at v_b:

$$Z_w = \frac{-Y_w^c(v_b\sin\theta - v_c\sin\theta + f\cos\theta)}{v_b\cos\theta - v_c\cos\theta - f\sin\theta}. \quad (4)$$

From equation 4 and equation 3 we can now solve for the human height Y_w^h in the scene, given the top and bottom position of a human detection in an image:

$$\begin{aligned} Y_w^h =\ & \frac{\frac{-Y_w^c(v_b\sin\theta - v_c\sin\theta + f\cos\theta)}{v_b\cos\theta - v_c\cos\theta - f\sin\theta}(v_t\cos\theta - v_c\cos\theta - f\sin\theta)}{f\cos\theta + v_t\sin\theta - v_c sin\theta} \\ & + \frac{Y_w^c(v_t\sin\theta - v_c\sin\theta + f\cos\theta)}{f\cos\theta + v_t\sin\theta - v_c sin\theta}. \end{aligned} \quad (5)$$
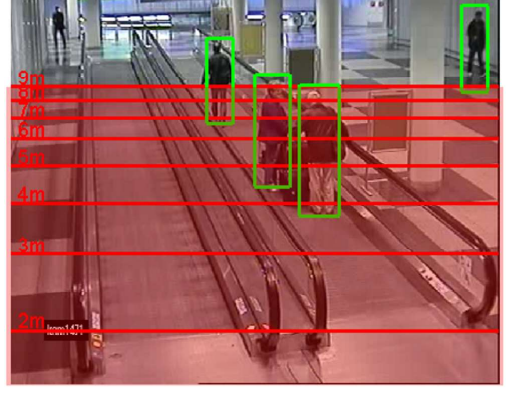

Figure 5. Visualization of the ground plane estimation for one camera.

4.2.2 Maximum Likelihood Estimation

When assuming the camera optical center to be in the center of the image plane, three camera parameters α for the simplified camera model remain: camera tilt θ, camera height Y_w^c and focal length f (cf. equation 1 and equation 2); camera yaw does not need to be estimated for the present task. Based on N reliable (i.e. highly scored) human detections $\omega = [(v_t, v_b)_1, \ldots, (v_t, v_b)_N]$ within a set of frames for a camera, we can determine the camera parameters $\alpha = [\theta, f, Y_w^c]$ by solving

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}}\ p(\omega|\alpha). \quad (6)$$

In order to estimate the camera parameter α, we assume that the heights of humans follow a Gaussian distribution $\mathcal{N}$ with known mean μ and variance σ. Taking the logarithm of equation 6, the optimal camera parameters $\hat{\alpha}$ align the distribution of reconstructed human heights $Y_{w\,i}^h$ with the human height prior p such that

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}}[\frac{N}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i^N (Y_{w\,i}^h(\alpha) - \mu)^2]. \quad (7)$$

This in turn allows computing metric heights for each object detection. Although the estimation results are approximate and depend strongly on the presence of a sufficient number of accurate human detections and on the accuracy of the human height prior, they are sufficient to discard detection outliers which are inconsistent with the estimated 3D scene and provide the framework for the subsequent 3D object-level context; in addition, the metric information can be used in the subsequent retrieval queries. See figure 5 for a visualization of the reconstructed ground plane for one camera.

4.3. 3D Object-Level Spatial Context

We aim at modeling the 3D spatial context for different types of luggage. Since the possibilities of interaction between a person and a given piece of luggage are limited and the availability of 3D scene information allows us to model
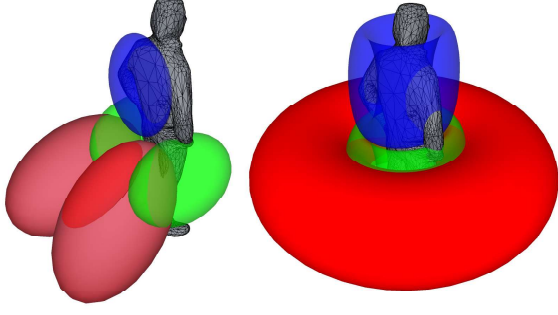
Figure 6. 3D spatial context for luggage carried by a person: if human pose information is provided by the detector, the probability of occurrence of different types of luggage (red: trolley, green: suitcase, blue: backpack) can be modeled with 3D Gaussian ellipsoids (left) whose 2D projections can be computed efficiently. In the simplified case of pose-free 2D human detections, the ellipsoids become rotation-invariant tori (right).

their co-occurrence in 3D, we propose a static 3D context model as shown on the left of figure 6. If we assume that the detection step provides an estimate of the person's orientation, the probability of occurrence of each type of luggage can be modeled with a 3D Gaussian $\mathcal{N}$ with center $\mu_{(3D)}$ and covariance $C_{(3D)}$ such that the probability of occurrence of a piece of luggage at a 3D position $\mathbf{X_w}$ follows

$$p(\mathbf{X_w}|v, \mu_{(3D)}, C_{(3D)}) = \mathcal{N}\left(\mathbf{X_w}|P_v \mu_{(3D)}, R_v C_{(3D)}\right) \quad (8)$$

where for a given human pose v, $P_v = [R_v|t_v]$ with rotation R_v and translation t_v. From the calibration step, the projection Φ_α for given camera parameters $\alpha = [\theta, f, Y_w^c]$ allows to derive the probability of occurrence at position $\mathbf{u}$ in image space from

$$p(\mathbf{u}|v, \mu_{(3D)}, C_{(3D)}, \alpha) = \mathcal{N}\left(\mathbf{u}|\Phi_\alpha(P_v \mu_{(3D)}), \Phi_\alpha(R_v C_{(3D)})\right). \quad (9)$$

To simplify the computation of the projected covariance, we approximate Φ_α by a Taylor expansion localized at $P_v \mu_{(3D)}$ and assume the projection to be locally affine,

$$\Phi_\alpha(R_v C_{(3D)}) \approx J_{\Phi_\alpha}(P_v \mu_{(3D)}) \cdot R_v C_{(3D)} \cdot J_{\Phi_\alpha}^t(P_v \mu_{(3D)}) \quad (10)$$

where J_{Φ_α} is the Jacobian of the projection Φ_α. If no orientation is provided as part of the human detection, the luggage co-occurrences are modeled as rotation-invariant tori centered around the human (figure 6, right) which can be approximated by a set of Gaussians whose projections can be derived analogously. As a result, luggage detections can be re-scored based on their probability of occurrence, given the nearby human detections and the 3D scene estimation.

4.4. Attribute Extraction

For each scored contextual detection pair of a human and a piece of luggage, we extract a set of semantic attributes

as described below and store the detections with their attributes in a database (see figure 3). The attributes can then be used to answer combined queries such as "tall person wearing a light-colored shirt and black pants, carrying a red trolley" in an efficient way. The query results are presented to the user in descending order based on the detection score of the piece of luggage and its contextual weight derived from equation 9.

Luggage type: The output of the object detector after contextual filtering for the three different pieces of luggage (trolley, backpack and suitcase) can be included in the query.

Luggage color: We extract the dominant color for each detected piece of luggage, following the approach of [10]: the HSL color space is quantized into 5 colors - red, green, blue, white and black. The dominant color for each luggage detection is computed by converting each detection into the HSL space and assigning each pixel to one of the five predefined colors. The color with the majority of votes is then assigned as the dominant color of the detection, which can be included in the user query. The color quantization has only limited robustness towards lighting changes, but it is sufficient to answer user queries containing relative color attributes such as light-colored, red, green, blue or dark.

Human height: Since each human detection is assigned a metric height from section 4.2.2, the user can specify height constraints for a human. Note that the queries can only be answered within the precision of the 3D reconstruction, which is approximate and may display slight variations for each set of frames processed as outlined in section 4.2.2. Consequently, approximate query formulations, e.g. "of small height", are preferable and correspond better to typical user input. We quantize human height attributes into the three categories small, average and tall, based on variation levels in the prior human height distribution.

Human color: We divide the area of each human detection box into an upper part and a lower part; each part is color-quantized separately in the same way as for the luggage. As a result queries such as "person wearing a light-colored shirt and blue pants" can be answered.

5. Experimental Results

In this section, we summarize the results achieved with a Matlab implementation of the present approach on a realistic data set. We describe the synthetic training data set, the test data set and the impact of the proposed context model on object class detection and query retrieval precision.

5.1. Data Set

From an airport surveillance network we collected 6148 frames from three cameras under different pan-tilt-zoom settings. Although the data set also contains some consecutive sequences, no overall temporal consistency of the

Figure 7. Filtered detections after each processing step (cf. figure 3). Initial 2D detections (left) of humans (green) and backpacks (red) show the difficulty of detecting inconspicuous objects. After 3D auto-calibration, detections which are inconsistent with scene scale and geometry can be removed (center), but different luggage types still cannot be differentiated reliably. When applying the 3D context model centered around the human detections, the precision of the backpack detector is significantly improved (right).
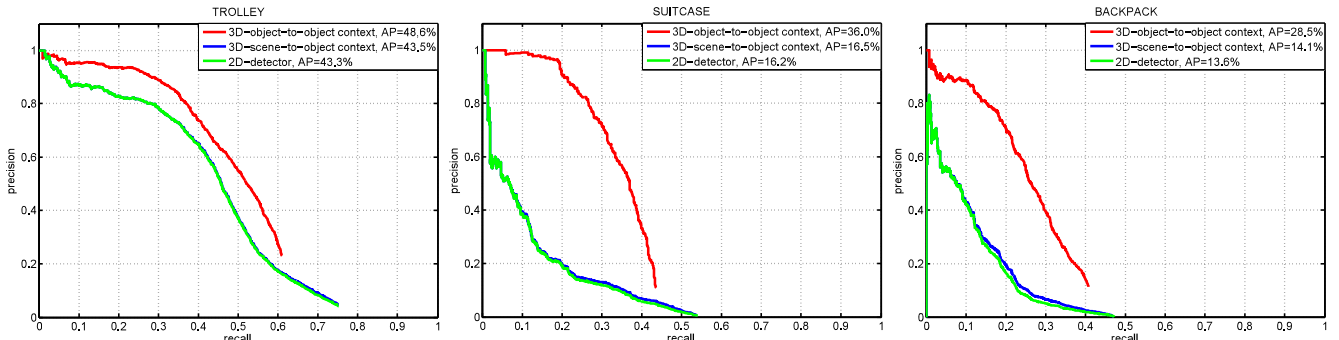

Figure 8. Impact of the proposed 3D context model on object class detector performance for the three luggage types trolley (left), suitcase (center) and backpack (right) on the entire test data set for all cameras. We show the pure 2D detector results (green), the results after filtering outliers based on the metric scene calibration (blue) and the final result of the 3D object-level context (red).

recorded frames is present. For evaluation the test data set was manually annotated with ground truth labels indicating image location and object class for three types of luggage: trolleys (3217 annotated object instances in the entire data set), suitcases (995 annotated instances), and backpacks (2051 annotated instances). For the reconstruction we assume a human height distribution with a mean of $1.8m$ and a standard deviation of $0.1m$.

5.2. Training Data

The object detectors for each object class were trained using synthetic 3D models. We purchased a number of textured CAD models from reseller *turbosquid.com* to account for the typical variations in object class appearance, notably 32 humans in different poses, 2 trolleys, 5 backpacks and 9 suitcases. The models were rendered in front of all real negative training images from the data set of [4] in addition to 30 images of typical airport background scenes not contained in the test set. Figure 4 shows some examples. Training annotations are automatically generated from the projected bounding boxes of the CAD models. Training follows the standard procedure outlined in [8].

5.3. Object Class Detection

We assess the impact of the proposed 3D object-level context model relative to the baseline performance of the individual 2D detectors before and after metrically filtering with the estimated 3D scene geometry: Figure 8 compares precision vs. recall for the three luggage classes obtained with the pure 2D detections (green), the 2D detections after metrically filtering based on the 3D scene reconstruction (blue) and the result when incorporating our 3D object-level context model (red). Since the 2D detector currently does not provide pose information, we use the rotation-invariant context model as outlined in section 4.3. The baseline detections for an example frame and the backpack detector are illustrated in figure 7 before (left) and after (center) metric filtering based on 3D scene context. The right image of figure 7 shows the output of the final 3D object-to-object context model. Note that the backpack class is the least discriminative of the three luggage classes, and the detector output is relatively unreliable. Even after removal of metrically impossible detections, its precision is insufficient for the intended retrieval task. In figure 8 we compare precision vs. recall for each object class detection evaluated on the entire test set. We observe a notable increase in aver-

age precision for all three luggage classes, ranging between $5\% - 20\%$. The trolley class is the most discriminative of the three classes; consequently, the performance gain is less pronounced. For the other two classes, the achieved gain in precision is crucial for the subsequent query tasks. Note that some recall is lost in the 3D context step, since the human detector fails to detect some context-relevant persons, usually due to significant occlusion. The detector performance is not significantly improved by metrically filtering based on the reconstructed scene constraints alone, since the amount of high-scoring false positive detections for inconspicuous object classes remains large.

5.4. Query

In order to evaluate the impact of the proposed approach on a retrieval task, we determined the retrieval precision at rank 25 ($P(25)$) for different queries of varying complexity for which we could guarantee that at least 25 query-relevant frames were present in the test set; this evaluation criterion is used in traditional image retrieval benchmarks[1]. Note that in the test data set, people can be captured multiple times by the same camera or appear in more than one camera. For the simple queries, we compare the retrieval precision of the 2D detectors and the context model; the complex queries involving metric information can only be answered using the context model, thus no comparison is given. Since the groundtruth annotations only indicate luggage location and type, no metadata is available to automatically evaluate complex color-based and metric queries; consequently, we determined $P(25)$ for our test queries manually. All retrieval results are ranked based on their scores. Figure 9 plots $P(25)$ for a few selected queries; figures 10,11,12 show some examples of retrieved frames for differently complex queries; we omit results showing the same person over several frames in order to illustrate the variation in the retrieval result. Although the color-quantization fails for some lighting configurations, our 3D context approach can significantly improve the retrieval results over the individual 2D detectors. Once again, the performance depends significantly on the discernibility of each luggage class and on the presence of reliable human detections; still, it is apparent that the query results would not be useable without the proposed 3D context model.

6. Conclusion

In the present work we describe an approach to retrieve frames from a network of surveillance cameras based on complex attribute-based queries for persons and the luggage they carry. We show that a simple 3D spatial context model in conjunction with an approximate 3D auto-calibration can significantly improve the performance of object class detectors for inconspicuous object classes such as different types
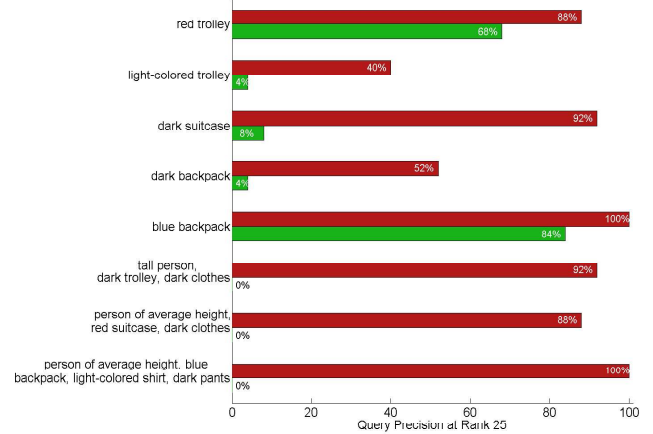
---

[1] http://www.imageclef.org

Figure 9. Improvements of precision at rank 25 for different queries when using our 3D object-level context (red). Note that complex queries cannot be answered with the 2D detections alone (green).
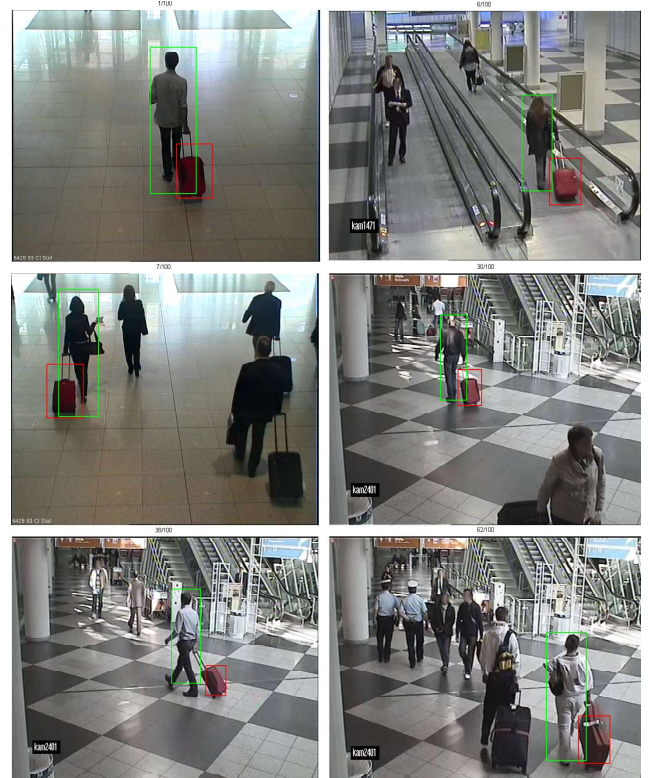

Figure 10. Example responses for the query "human of average height with red trolley".

of luggage, which might otherwise not be detectable individually. Future work will focus on integrating recent results on human pose estimation from single images in order to further improve the discriminatory power of the 3D context model.
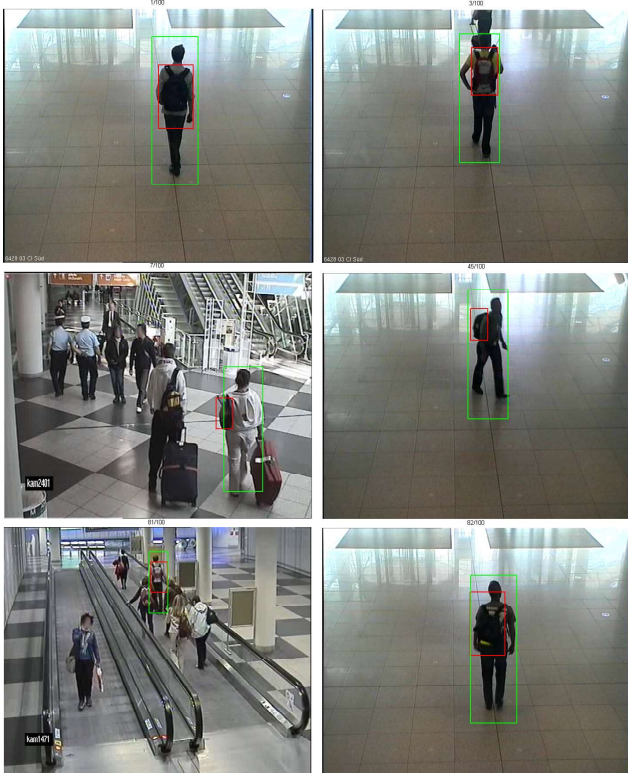
Figure 11. Example responses for the query "human of average height with dark backpack, light-colored shirt and dark pants".
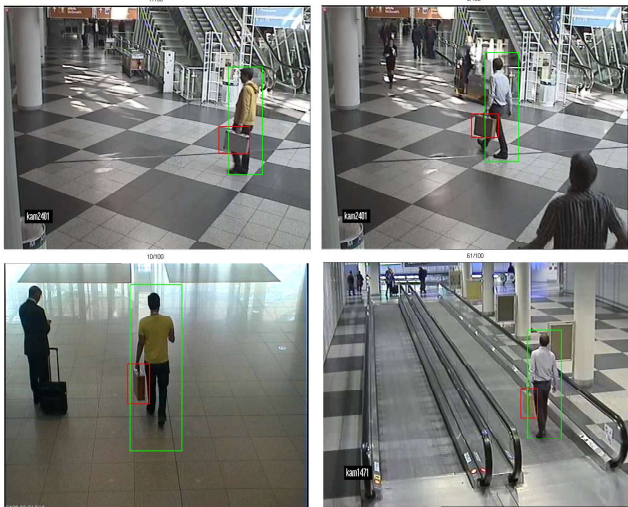

Figure 12. Example responses for the query "tall human with dark suitcase, light-colored shirt and dark pants".

References

[1] Y. Bao, M. Sun, and S. Savarese. Toward coherent object detection and scene layout understanding. In *CVPR*, 2010.

[2] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR*, 2011.

[3] M. Breitenstein, E. Sommerlade, B. Leibe, L. V. Gool, and I. Reid. Probabilistic parameter selection for learning scene structure from video. In *BMVC*, 2008.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[5] D. Damen and D. Hogg. Detecting carried objects in short video sequences. In *ECCV*, 2008.

[6] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, 2009.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results, 2010.

[8] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[9] R. Feris, J. Petterson, B. Siddiquie, L. Brown, and S. Pankanti. Large-scale vehicle detection in challenging urban surveillance environments. In *WACV*, 2011.

[10] R. Feris, B. Siddiquie, Y. Zhai, J. Petterson, L. Brown, and S. Pankanti. Attribute-based vehicle search in crowded surveillance videos. In *ICMR*, 2011.

[11] D. Hoiem, A. Efros, and M. Hebert. Putting objects into perspective. In *CVPR*, 2006.

[12] I. N. Junejo and H. Foroosh. Trajectory rectification and path modeling for video surveillance. In *ICCV*, 2007.

[13] J. Liebelt, C. Schmid, and K. Schertler. Viewpoint-independent object class detection using 3D feature maps. In *CVPR*, 2008.

[14] F. Lv, T. Zhao, and R. Nevatia. Self-calibration of a camera from video of a walking human. In *ICPR*, 2002.

[15] J. Marín, D. Vázquez, D. Gerónimo, and A. M. López. Learning appearance in virtual scenarios for pedestrian detection. In *CVPR*, 2010.

[16] B. Micusik and T. Pajdla. Simultaneous surveillance camera calibration and foot-head homology estimation from human detections. In *CVPR*, 2010.

[17] L. Pishchulin, A. Jain, C. Wojek, M. Andriluka, T. Thormählen, and B. Schiele. Learning people detection models from few training samples. In *CVPR*, 2011.

[18] J. Renno, J. Orwell, and G. Jones. Learning surveillance tracking models for the self-calibrated ground plane. In *BMVC*, 2002.

[19] J. Schels, J. Liebelt, K. Schertler, and R. Lienhart. Synthetically trained multi-view object class and viewpoint detection for advanced image retrieval. In *ICMR*, 2011.

[20] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR*, 2011.

[21] S. Stalder, H. Grabner, and L. V. Gool. Exploring context to learn scene specific object detectors. In *CVPR PETS workshop*, 2009.

[22] D. Tao, X. Li, X. Wu, and S. J. Maybank. Human carrying status in visual surveillance. In *ICCV*, 2006.

[23] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. In *WACV*, 2009.

[24] C. Wojek, S. Walk, S. Roth, and B. Schiele. Monocular 3D scene understanding with explicit occlusion reasoning. In *CVPR*, 2011.