People Detection in Fish-eye Top-views

Meltem Demirkus, Ling Wang, Michael Eschey, Herbert Kaestle and Fabio Galasso
Corporate Innovation, OSRAM GmbH, Munich, Germany

Keywords: People Detection, Top View, Fish Eye Lens, ACF, Grid of Classifiers.

Abstract: Is the detection of people in top views any easier than from the much researched canonical fronto-parallel views (e.g. Caltech and INRIA pedestrian datasets)? We show that in both cases people appearance variability and false positives in the background limit performance. Additionally, we demonstrate that the use of fish-eye lenses further complicates the top-view people detection, since the person viewpoint ranges from nearly-frontal, at the periphery of the image, to perfect top-views, in the image center, where only the head and shoulder top profiles are visible. We contribute a new top-view fish-eye benchmark, we experiment with a state-of-the-art person detector (ACF) and evaluate approaches which balance less variability of appearance (grid of classifiers) with the available amount of data for training. Our results indicate the importance of data abundance over the model complexity and additionally stress the importance of an exact geometric understanding of the problem, which we also contribute here.

1 INTRODUCTION

The detection of people has large relevance in the understanding of static and moving scenes, esp. since actions, human interactions and surveillance scenarios usually revolve around people. When possible, a top-view camera is preferable because it reduces the amount of person-person occlusion (cf. (Tang et al., 2014)). Additionally fish-eye lenses enlarge the field of view, which simplifies longer-term tracking, essential for the higher level understandings.

This work considers the nearly unaddressed topic of top-view person detection with fish-eye lenses from single images. Much research has addressed fronto-parallel views (Benenson et al., 2014), and there has been considerable work on surveillance scenarios (Stauffer and Grimson, 1999; Kaur and Singh, 2014; Rodriguez and Shah, 2007) but the frame-based top-view detection has only appeared recently (Chiang and Wang, 2014; Tasson et al., 2015), not yet at mainstream video conferences.

As for fronto-parallel views such as in Caltech (Dollár et al., 2009) and INRIA (Dalal and Triggs, 2005), top-view person detection is complicated by the high variability of people poses, e.g. standing, turning, crouching etc., and the false alarms from the background. Additionally, esp. in the case of fish-eye lenses, the view-point of the observed people changes hugely as soon as they move from the periphery of the image (where they are approximately fronto-parallel-looking) to the image center, where only the top of their head and shoulder profiles are visible, as illustrated in Figure 1.

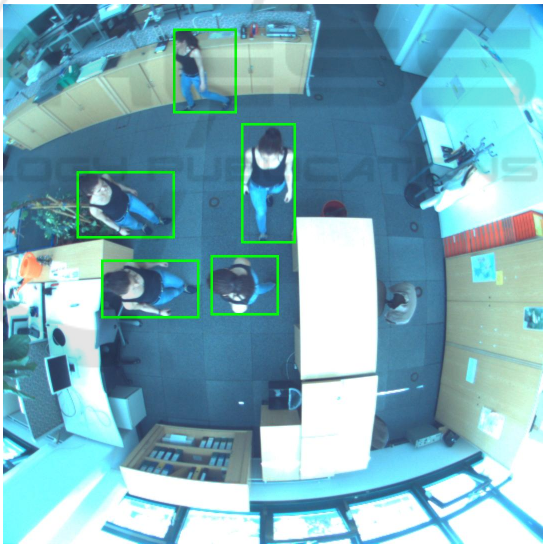


Figure 1: Detection of top-view people in fish-eye imagery is a challenging emerging topic. In addition to the difficulties of detection in canonical frontal views (pose changes and false detections), view-point changes dramatically complicates the task. We provide analysis of a state-of-the-art detection model, ACF (Dollar et al., 2014), geometric modelling of the fish-eye imagery and experimental evaluation on a new dataset.

Demirkus M., Wang L., Eschey M., Kaestle H. and Galasso F.
People Detection in Fish-eye Top-views.
DOI: 10.5220/0006094701410148
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 141-148
ISBN: 978-989-758-226-4

Here, we first acquire and annotate a dataset, large enough for model learning and evaluation, illustrated in Section 3. Then we consider the Aggregate Channel Feature (ACF) detector (Dollar et al., 2014) and extend it to grid of ACFs, to analyze the importance of the amount of training data Vs. the model complexity, in Section 4. We introduce the geometric model in Section 5, with an empirical study of its symmetry. Experiments in Section 6 support our statements, i.e. the open challenges, the importance of data and geometry.

2 RELATED WORK

Over decades of research, people detection literature has achieved grand results (Dollár et al., 2012; Benenson et al., 2014; Zhang et al., 2016), largely aided by more challenging and realistic datasets, such as INRIA (Dalal and Triggs, 2005), Caltech (Dollár et al., 2009) and KITTI (Geiger et al., 2012). Here we review the main people detection models, then compare our work to surveillance literature. Finally we illustrate related work on top-view people detection and the relevant geometric modelling.

As maintained in (Benenson et al., 2014), three main models have been fuelling pedestrian detection research: deformable parts models (DPM) (Felzenszwalb et al., 2010), boosting and integral channel features (ICF) (Dollár et al., 2009) and deep learning techniques based on convolutional neural networks (Hosang et al., 2015; Tian et al., 2015b).

DPM models attract as they cast the pedestrian detection problem as a structured one, i.e. detecting a person implies finding all of its composing parts, e.g. head, torso, limbs. While the model comes with the promise of resolving occlusion and change of view-points thanks to the locality of parts, it remains more cumbersome on the computational side, requiring detailed engineering to reach real-time performance (Sadeghi and Forsyth, 2014). Further to their large computational complexity, DPM models are currently not state-of-the-art, thus we do not consider them here.

Boosting and ICF (Dollár et al., 2009) models derive directly from the pioneering work of (Viola and Jones, 2004). The latest aggregation channel features (ACF) (Dollar et al., 2014) improve the image representation (ten channels involving histograms of oriented gradients, cf. HOG, gradient magnitudes and color) while simplifying the aggregation (a mere pooling over 2x2 pixel grids). The resulting boosting-based ACF model is capable of 30 fps speed and its performance still remains close to the state-of-the-art. Even more interestingly, current best performance builds directly on top of ACF (Hosang et al., 2015; Tian et al., 2015b; Zhang et al., 2015; Tian et al., 2015a; Yang et al., 2015; Cai et al., 2015; Zhang et al., 2016), which it enriches with CNN features. Both from a computational and performance perspective, ACF makes the most sensible candidate for our top-view person detection work. We extend ACF to include grids of ACF (cf. Section 4) and the geometric lens modelling (cf. Section 5).

Successful CNN models have so far built on top of ACF (Hosang et al., 2015; Tian et al., 2015b; Tian et al., 2015a; Cai et al., 2015) but have not replaced it. Additionally, CNN models require much larger computation and do not yet reach the efficiency of ACF. While the recent proposed model (Angelova et al., 2015) may achieve up to 15 fps, it still requires a GPU and gigabytes of memory. Since we aim to a deployment of our top-view detection system onto an embedded device, we omit consideration of CNN models here, proposing them as future extension of this work.

There is a large body of surveillance literature which relates to our work (Rodriguez et al., 2009; Roth et al., 2009; Rodriguez et al., 2011; Sternig et al., 2012; Corvee et al., 2012; Paul et al., 2013; Idrees et al., 2015; Solera et al., 2016). Surveillance generally implies fixed cameras detecting and tracking people from a viewing angle of $\sim 45^{\circ}$. Differently from this, our top-view imaging is acquired from a $\sim 90^{\circ}$ (zenithal) viewing angle. Our typical pedestrian appearance includes therefore the fronto-parallel and $\sim 45^{\circ}$ of surveillance (peripheral image areas), but it additionally includes top views (central areas), with only head and shoulders visible. Additionally, we adopt a fish-eye lens and, most importantly, we uniquely consider the detection task, as we believe this crucial for performance, before a tracking-based temporal reasoning.

Among the surveillance work, we draw inspiration from grid of classifiers (Roth et al., 2009; Sternig et al., 2012) and ask ourselves: may performance improve if different classifiers are adopted for diverse viewing poses? We experimentally answer this question in Section 5.

Two very recent papers are relevant to our work, i.e. (Chiang and Wang, 2014; Tasson et al., 2015). Both of them consider top-views of people, for the first time. The first considers a simple HOG+SVM (Dalal and Triggs, 2005), while the second adopts DPM. We draw inspiration from (Chiang and Wang, 2014) for initial geometric symmetry considerations and from (Tasson et al., 2015) for the labelling paradigm. We cannot compare to either work

as neither the code nor the datasets are available, but we are confident that a more state-of-the-art-aware choice of ACF would play to our advantage.

Finally, while previous work has studied the camera geometric modelling (Kannala and Brandt, 2006; Scaramuzza et al., 2006; Puig et al., 2012) for fish-eye lenses, to the best of our knowledge we are the first to employ a fish-eye geometric modelling in the context of people detection.

3 A NOVEL BENCHMARK FOR TOP-VIEW PEOPLE DETECTION WITH FISH-EYE LENSES

The novel benchmark should be large and challenging enough, to provide long lasting testbeds for new models and algorithms. We describe here the data acquisition, the non-trivial choice of an annotation standard and the proposed metrics.

3.1 Data Acquisition

As a valuable resource for training, validation and testing, the dataset should offer:

Person Imagery. The number of people in the dataset should be large enough and also, more importantly, there should not be overlaps between the training and test sets, as for generalization to unseen people;

Background Imagery. The trained person detector should generalize to new scenes. This requires therefore enough background imagery, also distinct across the training and test sets;

Repeatable Setup. Towards the future application and extension of this dataset for top-view activity recognition, tracking or social group formation, the installation should be repeatable. We set up therefore the camera facing down from the ceiling. We choose a commercial 2MP camera at 30 fps with a 200° field-of-view (FOV) lens. Altogether, this results in 1080x1080 images with 140° FOV.

The data collection is time-consuming, esp. the collection of background imagery as it implies reinstalling the camera in different physical locations. To maintain a balance between person/background samples, we take two kinds of data (sample frames are shown in Figure 2):

One-minute Videos with People. Footage where a person performs under the camera various poses in different positions, possibly interacting with the scene, e.g. sitting at a desk, moving objects. This is split into a test and training sets, maintaining the people and scene separately;

Figure 2: Sample dataset frames. Left and center images illustrate how videos of people were collected from any kind of background environment while the right image illustrates videos of empty scenes w/o person presence. The collected video data in top view perspective, which are generally unavailable in the common literature, enable background modelling with generalization to unknown scenes.

Single Frames of Background (BG). These images only contain empty scenes (for background samples) and are only used at training time (and do not overlap with the test set).

3.2 Data Annotation

It is an open research question how to label such data. Let us consider Figure 1. People walking at the peripheral areas of the image appear similar to the fronto-parallel views of Caltech (Dollár et al., 2009), with their head-feet axes directed towards the center of symmetry, approx. the center of the image. One might therefore choose a bounding box directed towards this center, as also done in (Chiang and Wang, 2014).

The annotation preference changes significantly as soon as the person moves towards the center of the image. There, the head and shoulder top-view profiles are only visible and the head-feet directional effect (towards the center) is negligible when compared to the body orientation. In the center, the people position in the image does not matter and a simple rectangular bounding box seems preferable, as chosen in (Tasson et al., 2015).

We follow up on the second direction for labelling: we ask the annotators to draw rectangular bounding boxes (axis-aligned) everywhere. We leverage therefore the labelling tool of (Dollár et al., 2009), which may interpolate annotated sparse frames across videos. We address and experiment on the alignment (crucial for training) in a post-processing step ((Drayer and Brox, 2014) similarly argues for machine alignment surpassing the user input accuracy).

3.3 Metrics

There is agreement in the literature (Benenson et al., 2014) on adopting the log-average miss rate

(LAMR) (Dollár et al., 2009). More specifically, a detection system takes in an image and finally return a list of detected BBs. The match of the detected BBs and the ground truth is rated by asserting an area overlap of the boxes of more than 50%. LAMR value then is defined as the geometric mean of detection miss rate values across different false positives per image (FPPI) rates, which finally provides a representative and robust estimate of the detection error at 0.1 false positives per image.

3.4 Dataset Description

Table 1 summarizes the dataset statistics. While each frame of the videos is labelled, for training and testing we adopt the same policy as in Caltech (Dollár et al., 2009) and only consider every 20th frame. This results in a total of 4459 selected labelled frames for training and 1736 selected labelled frames for testing.

Table 1: Dataset statistics at a glance.

	# Videos	# BG frames	# Labelled frames
Train set	37	84	89180
Test set	25	–	34720

4 ACF AND GRID OF ACFs FOR PEOPLE DETECTION

We consider for detection the Aggregate Channel Feature (ACF) (Dollar et al., 2014) due to performance (cf. (Benenson et al., 2014)) and availability, given the requirement of a CPU architecture. First we briefly introduce ACF, then we define a polar coordinate system to account for the near-circular symmetry.

4.1 The ACF Model

The ACF model adopts multi-scale multi-channel features in combination with a boosted tree classifier (cf. (Dollar et al., 2014) for details). Channels refer to the gradient magnitude, histograms of gradients and the image itself in the LUV color space. These are computed precisely over 4 octaves and are interpolated to 28 scales. Finally average pooling reduces the feature dimension by aggregation over 2x2 patches. Detection proceeds via classification of sliding windows, adopting shallow trees, boosted at learning time via hard-negative mining.

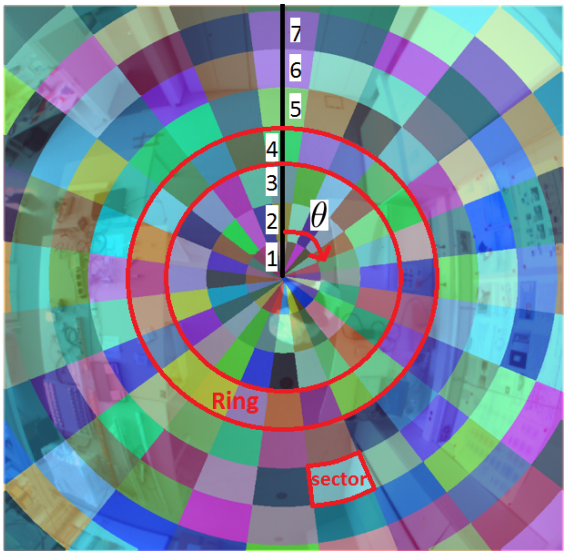


Figure 3: Acquired top-view images present an approximate circular symmetry. 7 rings are defined at increasing distances from the center (black numbers on the white text boxes). 24 sectors are defined (within rings) according to the polar angle θ, with origin at the vertical axis. Note the black vertical line: only people standing on it are *vertical* (heads north, feet south) as in most pedestrian datasets (Dollár et al., 2009).

4.2 The Image Symmetry and Coordinate System

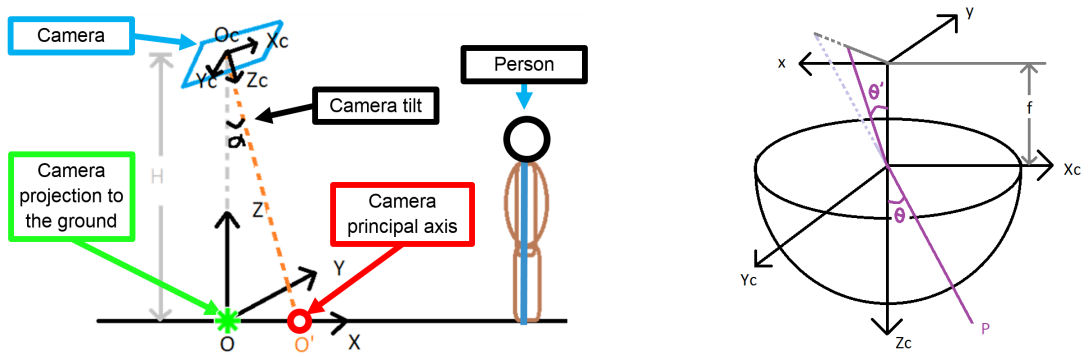
Let us refer again to Figure 1. The approximate circular symmetry of the image naturally favors the definition of rings and sectors for analysis. As illustrated in Figure 3, we define a system of polar coordinates based on rings (7 circular areas at increasing radial distances) and sectors within them (24 sectors, based on an angular coordinate θ from the vertical axis, each sized 15°). Walking from ring 7 to 1, people go from a frontal to a top-down view. Walking along θ (across sectors) they maintain their viewpoint but rotate.

5 TOP-VIEW FISH-EYE GEOMETRY

We define the system geometry and empirically study the effects of camera parameters and setup on its circular symmetry.

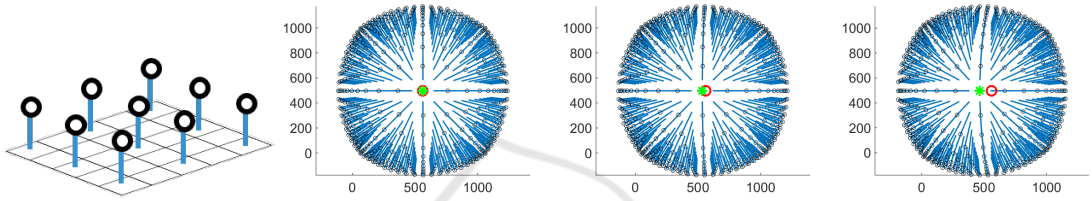
5.1 Setup and Fish-eye Imaging Model

Figure 4(a) presents a sketch of the setup, with a standing person observed by a ceiling mounted camera. World (homogeneous) coordinates **X** are defined



(a) Scene setup (b) Fish-eye lens

Figure 4: Scene setup (see Section 5.1) and fish-eye lens model (Kannala and Brandt, 2006).



(a) People setup (b) $\alpha = 0°$ (c) $\alpha = 3°$ (d) $\alpha = 10°$

Figure 5: Simulation of camera projection. Based on Equation (1) a matrix array of stick figures (*a*), representing standing people, are imaged (*b*,*c*,*d*) by a camera mounted with varying tilt angles α.

on the ground floor (independent from the camera mounting), camera coordinates $\mathbf{X_c}$ are consistent with the camera and α is the camera tilt, as the installed camera might not be perfectly vertical. We have:

$$\mathbf{u} = K\, G(\theta)\, [R|\mathbf{t}]\, \mathbf{X} \qquad (1)$$

with K the camera calibration matrix, R and $\mathbf{t}$ the world-to-camera rotation and translation, $\mathbf{u}$ the pixel coordinates (Hartley and Zisserman, 2004).

Equation (1) differs from the pin-hole camera by the matrix $G(\theta)$, describing the relation between the incoming and outgoing ray angles, cf. Figure 4(b). In more detail:

$$G(\theta) = \begin{pmatrix} (\tan \circ\, g)(\theta) & 0 & 0 \\ 0 & (\tan \circ\, g)(\theta) & 0 \\ 0 & 0 & \tan\theta \end{pmatrix} \qquad (2)$$

where function g models the radial distortion of the fish-eye lens. We calibrate the whole system as detailed in (Kannala and Brandt, 2006; Scaramuzza et al., 2006).

5.2 Quasi-circular Symmetry

Based on the calibrated projection model, we study the center of symmetry of the acquired images. To do so, we sketch standing people on a ground plane (Figure 5 (a)) and project them onto the images according to Equation (1) and different camera tilt angles α (Figure 5 (b,c,d)). The approximate center of symmetry is point O (*green* dot, re-projection of the camera *gravity* projection onto the ground floor) which shifts with the camera tilt α, thus with the camera installation angle. For small α's, point O' (*red* dot, projection of the camera principal axis) is close to O. Interestingly O' only depends on the camera calibration. We experiment on this in Section 6.3.

6 RESULTS AND DISCUSSION

We experiment on sample alignment, model complexity and finally on the effects of geometric modelling.

6.1 Training Sample Alignment by Rotation

First we address model aspect-ratio and the alignment of positive samples *for training*, two issues of importance for performance (Dollár et al., 2009; Benenson et al., 2014). For testing, we window-slide the computed models. The current bounding boxes (BB) are axis aligned, while the images depict people across 360° rotations (cf. Figure 1). We need therefore to rotate the samples to a reference angle and fit new BBs, tight on the person.

Table 2: Angle results: detection results of including samples from different angles (samples in ring 6 are used).

	LAMR in % (the lower the better)	
Selected sectors (in ring 6)	Subject-specific BB	Circumscribed BB
$-7.5^\circ < \theta < 7.5^\circ$	96.21	95.73
$-22.5^\circ < \theta < 22.5^\circ$	99.61	98.52
$-37.5^\circ < \theta < 37.5^\circ$	97.97	94.94
$-52.5^\circ < \theta < 52.5^\circ$	95.93	86.8

Table 3: ACF and Grid ACF results.

	LAMR in % (the lower the better)			
	Subject-specific BB		Circumscribed BB	
Selected rings	Single ACF	Grid ACF	Single ACF	Grid ACF
{6}	69.25	69.25	64.02	64.02
{6,5}	67	64.35	68.23	68.21
{6,5,4}	62.79	69.41	66.11	61.22
{6,5,4,3}	65.46	74.01	66.11	67.56
{6,5,4,3,2}	68.93	88.68	70.7	76.74
{6,5,4,3,2,1}	69.94	89.09	70.99	81.43

Initially, we consider for analysis ring 6 just (cf. Figure 3), which factors out the people size variation. We align all samples to the vertical north axis (black line in Figure 3) according to their BB centers. We fix the aspect-ratio to the average over all samples on the reference vertical axis. Then, there are two options for fitting tight new BBs to the rotated ones:

Alignment by Circumscribed BB. The rotation of an off-vertical-axis BB determines a diamond. The simplest way to generate a new BB is by circumscribing a rectangle.

Alignment by Subject-specific BB. We measure subject-specific BBs at the vertical axis and fit these to the rotated diamonds. (This excludes from training a few videos with subjects not crossing the vertical lines at ring 6.)

Another important parameter is about padding/stretching to adapt the computed BB to the estimated aspect-ratio, i.e. most commonly the shortest width/height is extended by sampling more background pixels (*context*), but stretching is also possible (ACF names this `squarify`, here we choose the one with best performance at training time).

In Table 2, we analyze increasing values of people rotation, starting from the sector at ring 6 on the vertical axis, $-7.5^\circ < \theta < 7.5^\circ$, up to the maximum $-57.5^\circ < \theta < 57.5^\circ$. Larger θ-ranges mean more data and better performance. Unexpectedly, the use of person specific information does not help, although the large error weakens the finding.

6.2 ACF Vs. Grid ACF

Next, we experiment with the large appearance changes of people with the distances from the center, extending the task from ring 6 to the whole image gradually. We compare:

Single ACF. One model is learnt from all data, i.e. from all selected rings. This exposes the one learnt classifier to highly multi-modal data distributions, i.e. people viewpoints;

Grid ACF. Separate models are learnt for separate rings. This simplifies the classifier task, but increases model complexity, thus requiring more data.

As illustrated in Table 3, both BB alignments attain similar performance. More interestingly, single ACF always outperforms Grid ACF. This indicates that data abundance is more important, which is also supported by the best overall performance (62.79%) for the selected rings {6,5,4}, i.e. most positive training samples within the limited test area.

6.3 Effects of Geometric Modelling

Finally, we question whether using the correct center of symmetry (point O in Figure 4(a)), instead of the image center (cf. the previous results) would improve performance.

Since O is not available (it requires computing the camera tilt at each installation), we first analyze if O' may substitute for it, i.e. the calibrated principal axis projection (cf. Section 5.1). We measure the discrepancy of the BB size when using O' instead of O (function of the camera tilt) and compare it to the noise in the labelling (variation of the user-labelled BBs over

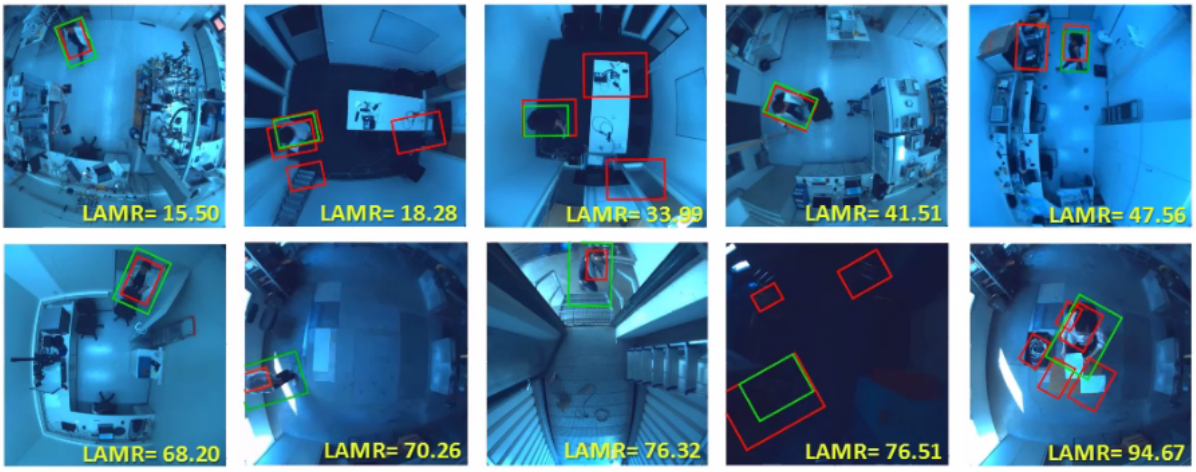


Figure 7: Sample detection results (*green* ground truth Vs. *red* detections), ordered per LAMR (best top-left, worst bottom-right). Background clutter and scarce illumination are major challenges.

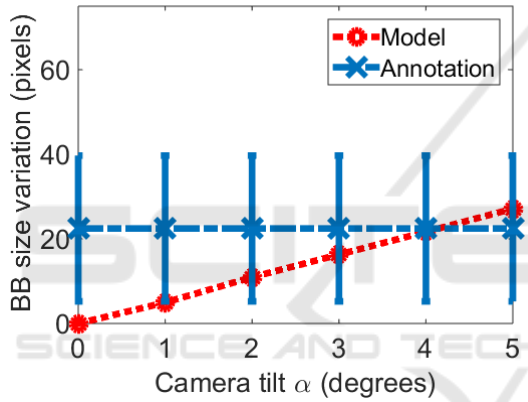


Figure 6: Center of symmetry approximation Vs. labelling noise.

the video, for the same person and image position). As shown in Figure 6, O' may approximate O up to the camera tilt of 4°. This applies to our dataset.

The approximate center of symmetry improves performance by 2% when using subject-specific BBs (67.99% LAMR, see Figure 7 for detection results), while the performance for circumscribed BBs reduces to 71.74%. Intuitively, the better modelling in geometry rewards the more accurate sampling alignment.

7 CONCLUSIONS

We have addressed pedestrian detection in top-views acquired with fish-eye optics. We have gathered a large dataset and analysed the importance of the annotation protocol with respect to the detection quality. Finally we have extended the state-of-the-art ACF detector to top-views by modelling the system geometry and found out that simpler models are preferable to richer grid ones, esp. if defining a grid implies reducing the amount of training data per model.

For the first time in top-view pedestrian detection, we have considered the background as a varying factor. By explicitly separating training and testing background imagery, we ensure that our detection results generalize across scenes and, most importantly, across scene variations, e.g. due to moving objects within it. To the best of our knowledge, this work considers the geometric modelling in relation to pedestrian detection for the first time. Interestingly, we have shown that geometry assumes more importance, the more the labelling can be accurately provided.

REFERENCES

Angelova, A., Krizhevsky, A., Vanhoucke, V., Ogale, A., and Ferguson, D. (2015). Real-time pedestrian detection with deep network cascades. In *BMVC*.

Benenson, R., Omran, M., Hosang, J., , and Schiele, B. (2014). Ten years of pedestrian detection, what have we learned? In *ECCV, CVRSUAD workshop*.

Cai, Z., Saberian, M., , and Vasconcelos, N. (2015). Learning complexity-aware cascades for deep pedestrian detection. In *ICCV*.

Chiang, A.-T. and Wang, Y. (2014). Human detection in fish-eye images using hog-based detectors over rotated windows. In *ICME Workshops*.

Corvee, E., Bak, S., and Bremond, F. (2012). People detection and re-identification for multi surveillance cameras. In *VISAPP - International Conference on Computer Vision Theory and Applications -2012*.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *CVPR*.

Dollar, P., Appel, R., Belongie, S., and Perona, P. (2014). Fast feature pyramids for object detection. *TPAMI*, 36(8):1532–1545.

Dollár, P., Tu, Z., Perona, P., and Belongie, S. (2009). Integral channel features. In *BMVC*.

Dollár, P., Wojek, C., Schiele, B., and Perona, P. (2009). Pedestrian detection: A benchmark. In *CVPR*.

Dollár, P., Wojek, C., Schiele, B., and Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *TPAMI*, 34.

Drayer, B. and Brox, T. (2014). Training deformable object models for human detection based on alignment and clustering. In *ECCV*.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *TPAMI*, 32(9):1627–1645.

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*.

Hartley, R. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press.

Hosang, J., Benenson, R., Omran, M., and Schiele, B. (2015). Taking a deeper look at pedestrians. In *CVPR*.

Idrees, H., Soomro, K., and Shah, M. (2015). Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(10):1986–1998.

Kannala, J. and Brandt, S. S. (2006). A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE Trans. Pattern Anal. Mach. Intell*, 28:1335–1340.

Kaur, R. and Singh, S. (2014). Background modelling, detection and tracking of human in video surveillance system. In *Computational Intelligence on Power, Energy and Controls with their impact on Humanity (CIPECH)*, pages 54–58.

Paul, M., Haque, S. M. E., and Chakraborty, S. (2013). Human detection in surveillance videos and its applications - a review. *EURASIP Journal on Advances in Signal Processing*, 2013(1):1–16.

Puig, L., Bermúdez, J., Sturm, P., and Guerrero, J. (2012). Calibration of omnidirectional cameras in practice: A comparison of methods. *Comput. Vis. Image Underst.*, 116(1):120–137.

Rodriguez, M., Ali, S., and Kanade, T. (2009). Tracking in unstructured crowded scenes. In *ICCV*.

Rodriguez, M., Sivic, J., Laptev, I., and Audibert, J.-Y. (2011). Density-aware person detection and tracking in crowds. In *ICCV*.

Rodriguez, M. D. and Shah, M. (2007). Detecting and segmenting humans in crowded scenes. In *ACM Multimedia*.

Roth, P., Sternig, S., Grabner, H., and Bischof, H. (2009). Classifier grids for robust adaptive object detection. In *CVPR*.

Sadeghi, M. A. and Forsyth, D. (2014). 30hz object detection with dpm v5. In *European Conference on Computer Vision*.

Scaramuzza, D., Martinelli, A., and Siegwart, R. (2006). A flexible technique for accurate omnidirectional camera calibration and structure from motion. In *International Conference on Computer Vision Systems (ICVS)*.

Solera, F., Calderara, S., and Cucchiara, R. (2016). Socially constrained structural learning for groups detection in crowd. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):995–1008.

Stauffer, C. and Grimson, W. (1999). Adaptive background mixture models for real-time tracking. In *CVPR*.

Sternig, S., Roth, P. M., and Bischof, H. (2012). On-line inverse multiple instance boosting for classifier grids. *Pattern Recogn. Lett.*, 33(7):890–897.

Tang, S., Andriluka, M., and Schiele, B. (2014). Detection and tracking of occluded people. *Int. J. Comput. Vision*.

Tasson, D., Montagnini, A., Marzotto, R., Farenzena, M., and Cristani, M. (2015). Fpga-based pedestrian detection under strong distortions. In *CVPR Workshops*.

Tian, Y., Luo, P., Wang, X., , and Tang, X. (2015a). Deep learning strong parts for pedestrian detection. In *ICCV*.

Tian, Y., Luo, P., Wang, X., and Tang, X. (2015b). Pedestrian detection aided by deep learning semantic tasks. In *CVPR*.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154.

Yang, B., Yan, J., Lei, Z., and Li, S. Z. (2015). Convolutional channel features. In *ICCV*.

Zhang, S., Benenson, R., Omran, M., Hosang, J., and Schiele, B. (2016). How far are we from solving pedestrian detection? In *CVPR*.

Zhang, S., Benenson, R., and Schiele, B. (2015). Filtered channel features for pedestrian detection. In *CVPR*.