Ground Truth For Pedestrian Analysis and Application to Camera Calibration

Clement Creusot
Toshiba R&D Center
Kawasaki, Japan
clementcreusot@gmail.com

Nicolas Courty
IRISA-UBS
Vannes, France
ncourty@gmail.com

Abstract

This paper investigates the use of synthetic 3D scenes to generate ground truth of pedestrian segmentation in 2D crowd video data. Manual segmentation of objects in videos is indeed one of the most time-consuming type of assisted labeling. A big gap in computer vision research can not be filled due to this lack of temporally dense and precise segmentation ground truth on large video samples. Such data is indeed essential to introduce machine learning techniques for automatic pedestrian segmentation, as well as many other applications involving occluded people. We present a new dataset of 1.8 million pedestrian silhouettes presenting human-to-human occlusion patterns likely to be seen in real crowd video data. To our knowledge, it is the first publicly available large dataset of pedestrian in crowd silhouettes. Solutions to generate and represent this data are detailed. We discuss ideas of how this ground truth can be used for a large number of computer vision applications and demonstrate it on a camera calibration toy problem.

1. Introduction

Computer vision problems related to scene understanding are so complex that they generally cannot be solved using human-designed heuristics. The main breakthroughs in last few decades has always been triggered by the introduction of new (or old) machine learning techniques to detect and recognize those objects. This kind of approach is not fully satisfying as the learning is often object-specific (*i.e.* it will only detect one object class: face, body, eye, car, and so on). None-the-less these techniques have been very efficient at solving a wide range of problems and have found many real-world applications in industry. One of the practical caveats with learning techniques is that a consequent amount of training data is required. In some cases it is possible to provide that ground-truth data manually or semi-automatically. This can be done either by a small group of people or by crowd sourcing (*e.g.* [2]) the labeling problem. Unfortunately, it is most of the time unpractical to use either of these "human-in-the-loop" techniques. Some of the main obstacles to produce human-supervised ground-truth are the lack of efficient semi-automatic labeling methods, training size requirements that make even simple monitoring/correction approaches too costly, and labeling quality requirements since intra and inter-human repeatability is not always tight enough.

Here we encounter a classic dead-lock problem where the data needed to design an automatic labeling system can not be reasonably obtained without an automatic labeling system in the first place. We propose to break that loop by using synthetic data to produce a first large set of ground truth silhouettes. We expect that it will help the development of better segmentation methods that in turn can be used to extract ground truth data from real video scenes with limited supervision.

In the next sections we detail the problem and related work on the subject before presenting the dataset and how it was generated. We then demonstrate on a toy problem how this data can be used to answer new questions and discuss the many possible applications this data can be used for.

2. Related Work

Most benchmarking videos used for pedestrian tracking offer no ground truth segmentation. The data usually provided consists of labeled rectangular bounding boxes around each pedestrian occurrence. A few datasets come with ground truth segmentation like [4] and [5], however the ground truth is only given for a few frames in the sequence. This is already praiseworthy as manually segmenting video is a very time-consuming task. It appears that no video dataset provides temporally dense segmentation. It has to be noted that there is no efficient tool to manually segment video. Some tools designed for tracking provide some contour detection functionality (*e.g.* [9]) but nothing that can be used for complex scenes.

In theory, 3D vision could really help the generation of ground truth segmentation for 2D videos (*e.g.* [3]). Segmenting objects or people using the depth information is indeed very easy. Once the segmentation is known, it can

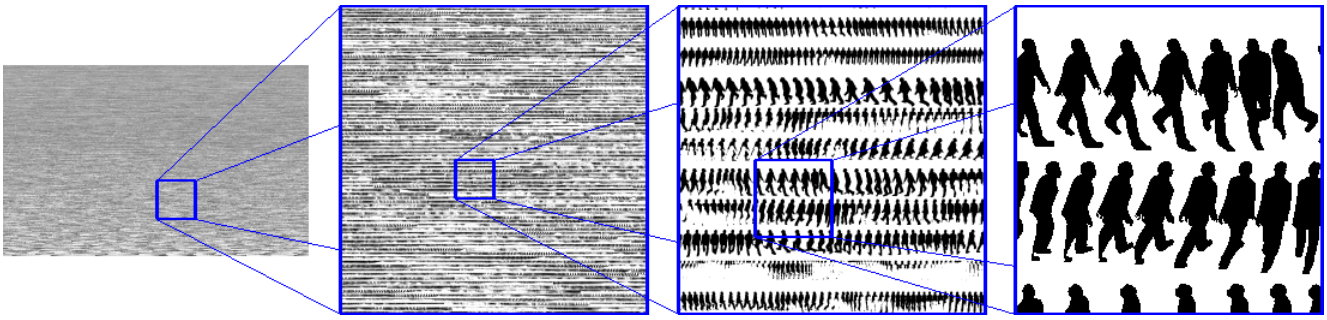


Figure 1. The dataset of 1.8 million pedestrian silhouettes can be visualized as a smooth 3.7 billions pixels pyramidal image. The right-most image is magnified 256 times.

be adjusted for the 2D color video. However it has not been used for pedestrian video datasets. Indeed, active 3D vision (*e.g.* structured light) camera usually have narrow reconstruction volumes and/or have trouble to capture outdoor scenes. Passive 3D vision reconstruction (*e.g.* stereo vision) is still difficult especially when dealing with uniform texture. In conclusion, there is no easy way to perform machine learning research on pedestrian segmentation for now and the main limiting reason is the lack of data. Using 3D synthetic human models to help computer vision applications is not a new idea. It was used in [10] for gait analysis applications, in [11] and [14] for action recognition using silhouettes, and in [13] to train the Microsoft Kinect body pose recognition system. To our knowledge, the dataset presented here is the first to provide individual pedestrian segmentation for crowded scenes and probably represents one of the largest human silhouette collections available to researchers.

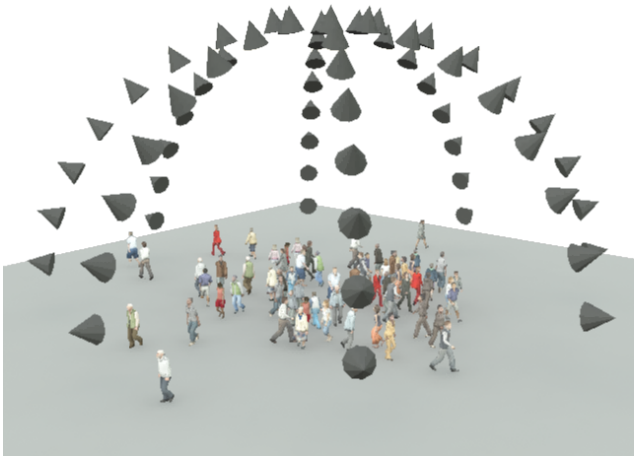


Figure 2. Simulated crossing scene along with the 64 camera configurations used to obtain the dataset.

3. Crowd Generation

Our dataset has followed the same guidelines for its creation as the recent synthetic crowd dataset AGORASET [1]. In this setting, 25 different avatars were used to produce an animation of four groups of 16 people going in opposite directions. In the middle of the crossing, a tangled pattern emerges as each pedestrian is trying to find his way through the crowd. The simulation model is a variation of Helbing's social force model [1] which guarantees that each pedestrian's path is collision free. Regarding the different avatars' animations, a walking motion capture file was used to drive the pedestrian skeletons. The playback speed for this motion was dynamically adjusted so as to match the current pedestrians velocities. The rendering was performed thanks to the commercially available Maya software. The rendering setup was adjusted so as to cancel any anti-aliasing filters that would have impaired the quality of the segmentation. As depicted in Figure 2, several cameras were positioned so as to render the different views of this scene. Using a simple polar parametrization, the azimuth and elevation angles distribution was discretized along a 8×8 grid. As a result, 250 frames were rendered through 64 cameras. In order to generate the silhouette, the scene is rendered with each pedestrian represented with plain colors. The colors are defined as a function of the pedestrian id in such way that any two pedestrian colors differ significantly and so that the inverse function is robust to noise. The inverse color function is used on the virtual camera output to produce 16-bit gray-level images where each pixel value corresponds to the pedestrian id. A 16-bit representation was chosen to allow more than 255 pedestrians in future scenes.

4. Dataset content

Extracting each individual silhouette from the global 16-bit mask is straightforward using thresholding filters on pixel values. The resulting image can be trimmed to a minimal bounding box and saved. Since all the parameters of the scene are known, we can also output for each pedestrian the

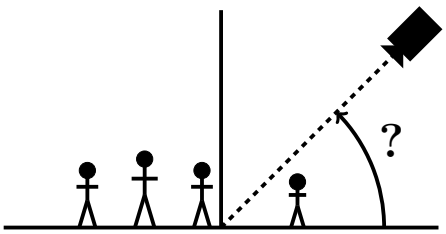


Figure 3. Toy problem: Is it possible to retrieve the camera tilt angle by solely using people shapes (N.B. not their size)?

number and identities of its occluders. To do that we consider that if two pixels associated with two different pedestrian touch each other, one of the pedestrian is occluded by the other. We select the occluder as being the one closer to the camera in that view using the absolute coordinates of the pedestrians (provided with the database). Pedestrian masks that are occluded by the image border get an additional occluder of label -1. This can be used to easily split the database into occluded and non occluded masks. The occluders' metadata is provided with the dataset.

From the 64 different views of 64 pedestrians in 250 different frames, we obtain 903,103 non-empty pedestrian masks. Since the mirror image of each mask is also a valid mask we end up with 1,806,206 individual silhouettes. In this dataset, 808,666 of the 1.8 million silhouettes are non-occluded. The complete set of silhouettes is shown as a single 3.7 billions pixels pyramidal image on the first author's website [1]. Figure 1 shows a sample zoom on the dataset.

In addition to the 16-bit images, a number of other ground-truth data is provided with the dataset. The bounding boxes for each pedestrian in each frame, their position, direction and speed in the global space referential, the skeleton data, the list of occluders for each pedestrian and for each camera, the association between avatar and pedestrian, the vertical angle for each pedestrian in each view (see second part of the article), as well as all camera parameters.

5. Practical Example: Retrieving the Camera Tilt Angle Using People Shapes

Is it possible to determine some camera parameters by only using the shape of the people in the video? In this section we present a toy problem where we try to find the tilt angle view α of one camera upon a crowded scene, even if only a single frame is given (see Figure 3). This proof of concept is designed to demonstrate the usefulness of our proposed ground-truth data to allow existing problems to be solved with original data-driven approaches.

[1] http://clementcreusot.com/pedestrian/

5.1. Simplification

The toy problem presented here only focuses on one camera calibration parameter (the tilt angle of the camera relative to the floor) to make the article more accessible to non-specialists. For example, determining the distance of the camera to the ground cannot be done without measuring something in the scene. This can be achieved by using a distribution of human heights or other techniques but it has no direct interest for our demonstration. Similarly the roll of the camera around his axis can be determined for low angle views by detecting vertical direction in the scene (like upright humans). As a first simplification, we will assume in the remainder that the camera is upright.

Camera parameters are divided in two categories: Intrinsic (like the focal length, principal point and image format(dimension of the sensor area) and extrinsic (like the position and orientation relative to other referentials). Here we assume an ideal pinhole camera model since distortions would not in any case matter for the type of scene and precision we are targeting. In practice, the intrinsic parameters do not cause many difficulties since most of them do not change over time (like the sensor size). In most cases, only the focal length varies, *e.g*. for zooming cameras. In this proof of concept we consider the focal length to be fixed. Note that when the view angle is retrieved, the focal length can be approximated using the formula in Figure 6 over a large set of silhouettes. This is not discussed here.

The main issue in uncontrolled environments are the extrinsic parameters. The position and orientation of the sensors is difficult to determine in large scale environment and, also, might change over time. Here the camera angle retrieved is relative to the ground floor and is positive, *i.e*. we consider the tilt angle of the camera rather than its pitch.

Note that in real life situations, getting the pedestrians' silhouettes using a simple 2D camera is very difficult. Here we first evaluate what the calibration performance would be if the segmentation problem was solved.

5.2. Rapid background

The idea of calibrating cameras using pedestrians is not new. Indeed, in most cases the aim of camera networks is to monitor people. Trying to use pedestrians as calibration devices is therefore a natural endeavor. Usually, calibration using pedestrians is done using the known or assumed heights of individuals in the picture [6][8][7]. In [6], the main assumption made is that the height of one person is stable throughout the gait motion. If the floor is flat and the camera still, you can draw the projection of parallel lines (in the 3D scene) by joining respectively the top head points and the bottom foot points of a single subject between frame p and q. Repeating this allows to get vanishing points that in turn can be used to determine the camera parameters. This foot-head homology method is interesting but can only be

α	10	20	30	40	50	60	70	80
α_y	13.6	27.3	40.4	53.0	64.7	75.8	86.4	96.9

Figure 4. Variations in pedestrian shape at position (x, y) w.r.t. the camera tilt angle α. Note that extracting a foot-head vector for high tilt angle view is impossible using naive top/bottom pixels selection.

used in very restricted conditions. First you need to track the individual within the video to be able to create those parallel lines and this can be a challenging task. Second, you need to determine the segment representing the person's height (top and bottom point), this is possible if the view angle is low but become impossible when the view tilt angle reaches 60 or above (see Figure 4).

In [7], cyclical inferences between object detections (cars and pedestrians) and the scene geometry are used to refine the camera position hypothesis for low tilt angle views.

In [8], the vertical vanishing point is computed using a RANdom SAmple Consensus (RANSAC) on vertical blob direction candidates extracted from a large number of pedestrians. The discovered vertical directions are then used to detect the head and foot points in each blob to compute the horizontal plane by a least mean square regression. The focal length is determine using hypothesis testing based on the known human heights distribution.

To our knowledge nobody has ever tried to learn the viewing angle using the shape of people's silhouettes. The closest related idea was found in [12], where the authors proposed to retrieve the camera angle using the ratio between the width and height of the pedestrian bounding box. However they did not use it in their final results.

Using people as measurement devices is not likely to work for a single person on a single frame but accumulation of clues from different people over time can give good approximations.

5.3. Setup

For our proof of concept the silhouette dataset was split into two disjoint parts. The training set is composed of 42 pedestrians (Ids 0-13, 25-38 and 50-63) corresponding to 14 avatars (0-13). The test set is composed of 22 pedestrians (Ids 14-24 and 39-49) corresponding to 11 different avatars (14-24). Figure 5 shows how the training and testing set are constructed.

This experiment is designed with video surveillance applications in mind, we therefore assume that the camera looks down on the scene, *i.e.* the world projection of the camera principal point is located under the horizon (so that the camera center axis intersect the ground plane).

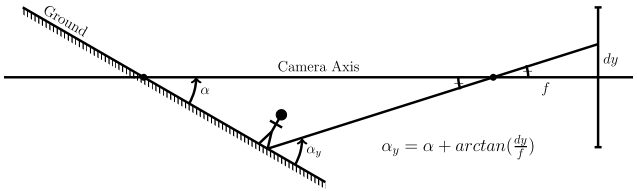


Figure 6. During training, the angle α_y is stored for each silhouette in the feature space.

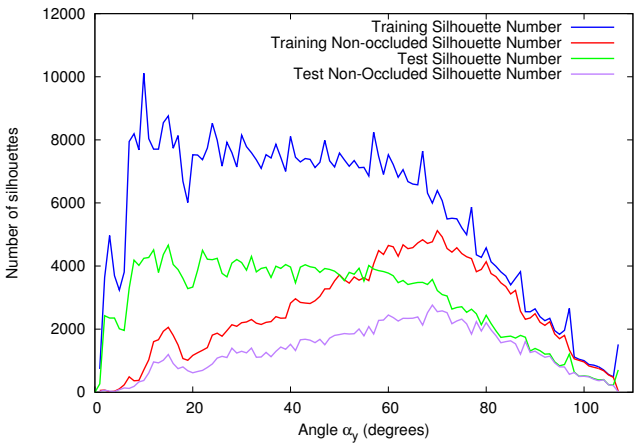


Figure 7. Distribution of α_y angles in the train and test sets.

In the offline part of the method, all training sequence masks are extracted to compute their representation feature vector and the relative view angle at which they are seen in the scene (see Figure 6). This angle can only be computed for a point, not for a whole object. The center bottom pixel of the pedestrian bounding box is used as a reference to determine α_y. The distribution of these angles in the train and test sets are given in Figure 7.

In the online part the input is a non-empty sequence of frames. For all the silhouettes in the sequence above a threshold area, feature vectors are computed. The view angle associated to the nearest neighbor in the feature space is retrieved. From this points two approaches are considered:

1. Simply averaging the angle returned by the matching process (we eliminate outliers using the 3-sigma rule).

2. Using all the angles associated with all the positions of the silhouettes in one frame to compute a regression of the view angle plane. Here a simple least square regression is performed. The angle for the center pixel of the camera is returned as the view angle for the frame.

Figure 5. Each scene view is split in two according to the pedestrian avatar group: train or test. This guaranty that all testing silhouettes represent previously unseen pedestrians.

(a) Shape Descriptors (b) 3 Nearest Neighbors (8x16)

Figure 8. Simple descriptors for shape matching.

The second approach requires all pedestrians to navigate on the same horizontal plane and assume the arctangent function in the α_y computation can be approximated by a linear function. This is true when the focal length is larger than half the sensor height. For example with a 35mm focal and a 36mm sensor, the function varies between $arctan(-0.51)$ and $arctan(0.51)$ and is almost a straight line. Results are given for all scenes with view angles between 10 and 80 degrees vertical angle for 100 frames (200 to 300). This represents 6400 different test images.

The features used to retrieve the angle are very coarse, the bounding box of the pedestrian silhouette is simply resized to fit a 16x32 (feature 1) or a 8x16 (feature 2) matrix which is serialized to a single feature vector. The matching is done by retrieving the first nearest neighbor in the feature space (see Figure 8). In the next Section we investigate how the results vary depending on several parameters of the system.

5.4. Results

For the toy problem presented here only synthetic video sequences have been considered. In order to test this idea on real data we would need both camera calibration ground-truth and dense segmentation ground-truth on the same sequences.

Prediction-combination strategies In Figure 9, we show the angle estimation errors obtained with our two approaches. A surprising fact is that even by simply averaging the angle estimation for each pedestrian in the scene we obtain errors under 10 degrees for low angles view and under 4 degrees for high tilt views. Please note that this is achieved by looking at a single frame only. Using a plane regression to retrieve the camera tilt angle seems beneficial for all angle views, lowering the error to under 7 degrees for low angle view and under 3 degrees for high tilt angles.

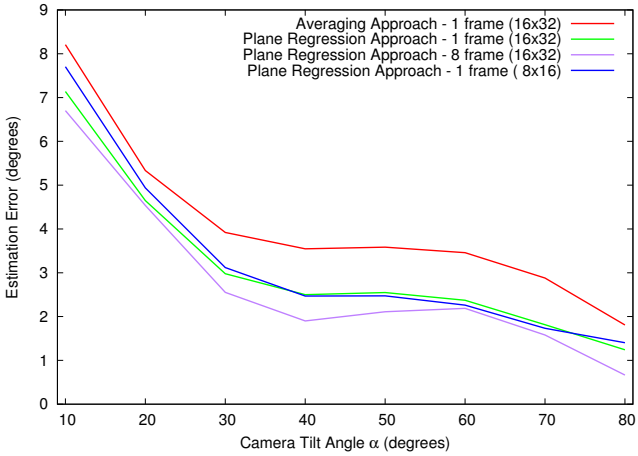


Figure 9. Influence of the computation approach, the number of frames and feature choice.

The difference of accuracy between low and high angle views seems to come from the shape matching approach. In Figure 10, we show the number of test silhouettes per frame. It can be seen that the total number of silhouettes is smaller for lower angles. Indeed, people in the foreground sometimes completely occlude people in the background. For the same reason the number of non-occluded masks is very low. Our shape matching mechanism does not know whether a shape is occluded or not, it just finds the closest looking shape seen in training. We think that developing an occlusion-aware shape comparison system might reduced the performance gap between low and high angle views.

Number of frames per prediction In Figure 9, we also compare the mean error according to the number of frames

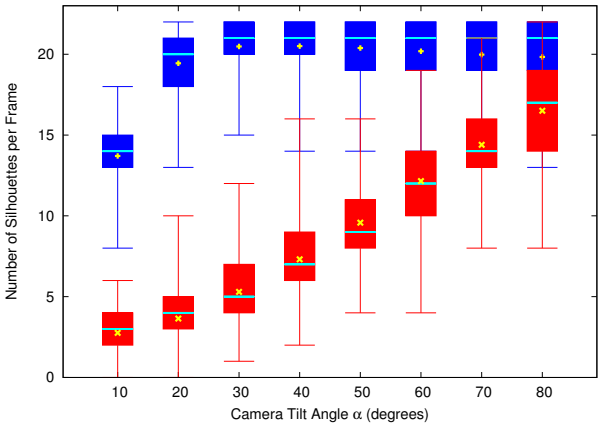


Figure 10. Test silhouette number per frame. The red distribution shows the number of non-occluded silhouettes, the blue one, the total number of silhouettes visible. Henceforth candle-sticks represent the min/Q1/med/Q3/max values, a yellow cross, the mean.

used for the angle estimation. As expected using more frames lowers the mean error but also reduces the deviation and worse case errors (maximum angle). Surprisingly however, the errors do not seem to converge to zero when the number of frames increases, as seen in Figure 11. These systematic errors toward which they converge seem to be small, but stable. The fact the error is constant for a given view angle is very interesting. We have not found any valid explanations for such behavior yet. It might be due to the fact we use the bottom of the bounding box to compute the angle α_y. This approximation has more impact on lower angle views.

In Figure 12 we show some visualization results obtained with our technique and a known focal length. The camera angle is used to create a virtual ground plane below the foreground silhouettes. The quality of the ground reconstruction and variation over frames are better seen in video (See supplemental material - `http://clementcreusot.com/pedestrian/`). While answering an interesting theoretical question about pedestrian shape, this work cannot yet be used on real life sequences as the people segmentation it requires is very difficult to obtain with current vision systems.

6. Discussion

In the field of segmentation, prior data is rarely considered. Most segmentation techniques are heuristic in nature or rely on very local learning (for example pixel color distributions). The contour of objects being often sufficient to recognize objects/postures, one might consider using silhouettes to help segmentation using machine learning techniques. This is our main focus for future work using this dataset.

We can imagine numerous other applications for which this data can be used. For example it might be possible to retrieve a person's orientation from its silhouette which might be of interest in terms of attention detection, urban planing, and advertisement study. Detecting if a silhouette is occluded or not can also be done by using two-fold classification methods. If a statistical model is constructed from the database it is in theory possible to reconstruct an estimate of the occluded part of a pedestrian. This can be very useful in highly crowded scenes where the level of occlusion is important.

While our ground-truth dataset of silhouettes can still be improved in terms of variability (limited number of avatars, gait, situations, lack of accessories and non-human occlusions) it is much more precise than any manually acquired pedestrian video segmentation. The strong point of our dataset is that it is temporally dense, pixel accurate, that it presents a variability large enough to generalized to previously unseen people, and offers a large amount of human-to-human occlusions.

References

[1] P. Allain, N. Courty, and T. Corpetti. Agoraset: a dataset for crowd video analysis. In *Proceedings of the 1st International Workshop on Pattern Recognition and Crowd Analysis*, 2012.

[2] Amazon. Mechanical turk. `http://www.mturk.com`.

[3] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV (1)*, pages 44–57, 2008.

[4] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *Proceedings of the 11th European conference on Computer vision: Part V*, ECCV'10, pages 282–295, Berlin, Heidelberg, 2010. Springer-Verlag.

[5] K. Fragkiadaki, W. Zhang, G. Zhang, and J. Shi. Two-granularity tracking: Mediating trajectory and detection graphs for tracking under occlusions. In *ECCV (5)*, pages 552–565, 2012.

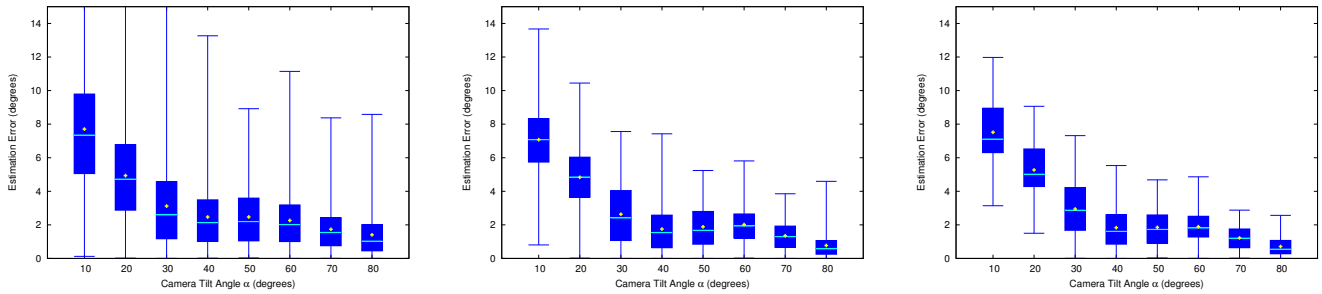
[6] M. Hödlmoser and M. Kampel. Multiple camera self-calibration and 3d reconstruction using pedestrians. In *ISVC (2)*, pages 1–10, 2010.

[7] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. *International Journal of Computer Vision*, 80(1):3–15, 2008.

[8] R. C. Jingchen Liu and Y. Liu. Automatic surveillance camera calibration without pedestrian tracking. In *Proceedings of the British Machine Vision Conference*, pages 117.1–117.11. BMVA Press, 2011.

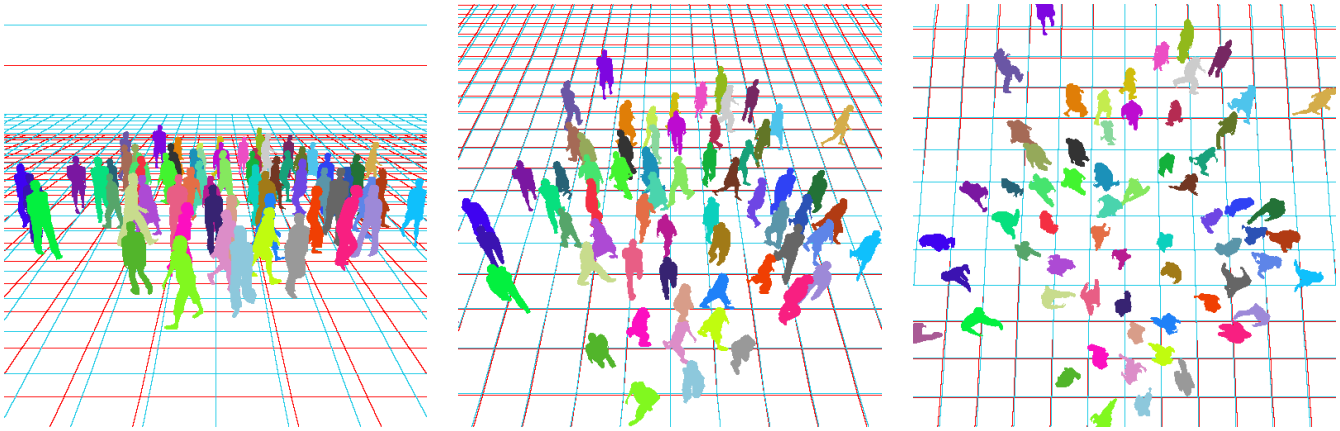
[9] I. Kavasidis, S. Palazzo, R. Di Salvo, D. Giordano, and C. Spampinato. A semi-automatic tool for detection and tracking ground truth generation in videos. In *Proceedings of the 1st International Workshop on Visual Interfaces for Ground Truth Collection in Computer Vision Applications*, VIGTA '12, 2012.

[10] Y. Liu, R. Collins, and Y. Tsin. Gait sequence analysis using frieze patterns. In A. Heyden, G. Sparr, M. Nielsen, and



a - 1 frame b - 8 frames c - 32 frames

Figure 11. Distribution of results obtained with the plane regression approach (8x16 features) using different numbers of frames. We notice a quasi-constant systematic error for each angle view.



20 degrees
Large Error: +4.8°

50 degrees
Small Error: -0.8°

80 degrees
Small Error: +1.1°

Figure 12. Example of virtual planes retrieved using our technique on a single frame using only 22 of the 64 shown pedestrians (test set). The ground truth is show in red, our result, in light blue. The frame-to-frame variation can only be seen in video (see supplemental material - `http://clementcreusot.com/pedestrian/`).

P. Johansen, editors, *Computer Vision ECCV 2002*, volume 2351 of *Lecture Notes in Computer Science*, pages 657–671. Springer Berlin Heidelberg, 2002.

[11] H. Ragheb, S. Velastin, P. Remagnino, and T. Ellis. Vihasi: Virtual human action silhouette data for the performance evaluation of silhouette-based action recognition methods. In *Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference on*, pages 1–10, 2008.

[12] D. Rother, K. Patwardhan, and G. Sapiro. What can casual walkers tell us about a 3d scene? In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, 2007.

[13] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-Time Human Pose Recognition in Parts from Single Depth Images. *Computer Vision and Pattern Recognition*, June 2011.

[14] S. Singh, S. Velastin, and H. Ragheb. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 48–55, 2010.