Exploiting Projective Geometry for View-Invariant Monocular Human Motion Analysis in Man-made Environments

Grégory Rogez[a,b], Carlos Orrite[a], J. J. Guerrero[a], Philip H. S. Torr[b]

[a] *Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, SPAIN*

[b] *Department of Computing, Oxford Brookes University, Oxford, UK*

Abstract

Example-based approaches have been very successful for human motion analysis but their accuracy strongly depends on the similarity of the viewpoint in testing and training images. In practice, roof-top cameras are widely used for video surveillance and are usually placed at a significant angle from the floor, which is different from typical training viewpoints. We present a methodology for view-invariant monocular human motion analysis in man-made environments in which we exploit some properties of projective geometry and the presence of numerous easy-to-detect straight lines. We also assume that observed people move on a known ground plane. First, we model body poses and silhouettes using a reduced set of training views. Then, during the online stage, the homography that relates the selected training plane to the input image points is calculated using the dominant 3D directions of the scene, the location on the ground plane and the camera view in both training and testing images. This homographic transformation is used to compensate for the changes in silhouette due to the novel viewpoint. In our experiments, we show that it can be employed in a bottom-up manner to align the input image to the training plane and process it with the corresponding view-based silhouette model, or top-down to project a candidate silhouette and match it in the image. We present qualitative and quantitative results on the CAVIAR dataset using both bottom-up and top-down types of framework and demonstrate the significant improvements of the proposed homographic alignment over a commonly used similarity transform.

Keywords: Human motion analysis, Projective geometry, View-invariance, Video-surveillance

1. Introduction

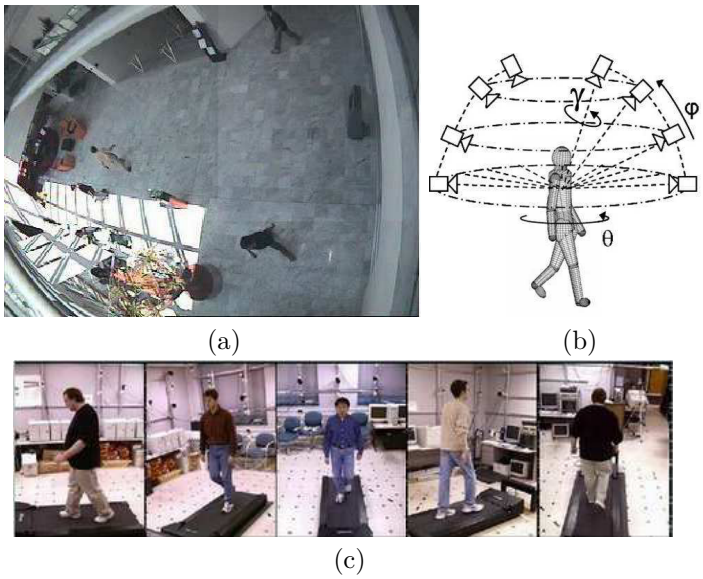
In recent years, the number of cameras deployed for surveillance and safety in urban environments has increased considerably in part due to their falling cost. The potential benefit of an automatic video understanding system in surveillance applications has stimulated much research in computer vision, especially in the areas related to human motion analysis. The hope is that an automatic video understanding system would enable a single operator to monitor many cameras over wide areas more reliably.

Example-based approaches have been very successful in the different stages of human motion analysis: detection, pose estimation and tracking. Some consist of comparing the observed image with a data base of stored samples as in [1, 2, 3]. In some other cases, the training examples are used to learn a mapping between image feature space and 3D pose space [4, 5, 6, 7]. Such mappings can be used in a *bottom-up* discriminative way [8] to directly infer a pose from an appearance descriptor or in a *top-down* generative manner [7] through a framework (e.g., a particle filter) where pose hypotheses are made and their appearances aligned with the image to evaluate the corresponding observation likelihood or cost function. The exemplars can also be used to train binary human detectors [9, 10, 11, 12], multi-class pose classifiers [13, 14, 15] or part-based detectors [16, 17, 18, 19, 20] that are later employed to scan images. The main disadvantage of all these example-based techniques is their direct dependence on the point of view: the accuracy of the result strongly depends on the similarity of the camera viewpoint between testing and training images.

Ideally, to deal with viewpoint dependency, one could generate training data from infinitely many camera viewpoints, ensuring that any possible camera viewpoint could be handled. Unfortunately, this set-up is physically impossible and makes the use of real data infeasible. It could, however, be simulated by using synthetic data, but using a large number of views would drastically increase the size of the training data. This would make the analysis much more complicated; furthermore, the problem is exacerbated when considering more actions.

In practice, roof-top cameras are widely used for video surveillance applications and are usually placed at a significant angle from the floor (see Fig. 1a), which is different from typical training viewpoints as shown in the examples in Fig. 1c. Perspective effects can deform the human appearance (e.g., silhouette features) in ways that prevent traditional techniques from being applied correctly. Freeing algorithms from the viewpoint dependency and solv-

ing the problem of perspective deformations is an urgent requirement for further practical applications in video-surveillance.



(a) (b)

(c)

Figure 1: (a) Testing video-surveillance video from Caviar database [21]. (b) Viewing hemisphere: the position of the camera with respect to the observed subject (the view) can be parameterized as the combination of two angles: the *elevation* $\varphi \in \left[0, \frac{\pi}{2}\right]$ (also called latitude or tilt angle) and the *azimuth* $\theta \in [-\pi, \pi]$ (also called longitude). A third angle $\gamma \in [-\pi, \pi]$ can be considered to parameterize the rotation around the viewing axis. (d) Examples of training images from the MoBo dataset [22].

The goal of this work is to track and estimate the pose of multiple people independently of the point of view from which the scene is observed (see Fig. 1b), even in cases of high tilt angles and perspective distortion. The idea is to model body pose manifold and image features (e.g., shape) using as few training views as possible. The challenge is then to make use of these models successfully on any possible sequence taken from a single fixed camera with an arbitrary viewing angle. A solution is proposed to the paradigm of *"View-insensitive process using view-based tools"* for video-surveillance applications in man-made environments: supposing that the observed person walks on a planar ground in a calibrated environment, we propose to compute the homography relating the image points to the training plane of the selected viewpoint. This homographic transformation can potentially be used in a bottom-up manner to align the input image to the training plane and process it with the corresponding view-based model, or top-down to project a candidate silhouette and match it in the image. In the presented work, we focus on specific motion sequences (walking), although our algorithm can be generalized for any action.

1.1. Related Work

Viewpoint dependence has been one of the bottlenecks for research development of human motion analysis as indicated in a recent survey [23]. Most of the early surveillance systems which can be found in the literature, eg W^4 [24], BraMBLe [25] or ADVISOR [26], only considered data where multiple people were distributed horizontally in the image, i.e., with a camera axis parallel to the ground and without any type of distortion. More recently, some work has focused on the problem of viewpoint dependency.

There have been successful efforts to build view-invariant features. The approach proposed in [27] exploits the invariances of Hu moments and the concept of "virtual cameras"which allows for the reconstruction of synthetic 2D features from any camera location. In [28], a calibrated approach was used in order to avoid perspective distortion of the extracted features while a method was proposed in [29] to build features that are highly stable under change of camera viewpoint and recognize action from new views.

Methods using ideas from model based invariance theory have been gaining popularity in recent years. In [30], the authors presented a method to calculate the 3D positions of various body landmarks given an uncalibrated perspective image and point correspondences in the image of the body landmarks. They also addressed the problem of view-invariance for action recognition in [31]. Recently, a motion estimation algorithm for projective cameras explicitly enforced articulation constraints and presented pose tracking results for binocular sequences [32]. In [33], the authors proposed a reconstruction method to rectify and normalize gait features recorded from different viewpoints into the side-view plane, exploiting such data for human recognition. The rectification method was based on the anthropometric properties of human limbs and the characteristics of the gait action [34].

The problem of comparing an input image (or a sequence of images) with a database of stored training exemples captured from different camera views has also been studied. Recently, view-invariant action recognition was achieved in [35] by associating a few motion capture examples using a novel Dynamic Manifold Warping (DMW) alignment algorithm. A solution for inferring a 3D shape from a single input silhouette with an unknown camera viewpoint was proposed in [36]: the model was learnt by collecting multi-view silhouette examples from a calibrated camera ring and the visual hull inference consisted in finding the shape hypotheses most likely to have generated the observed 2D contour. We also collect silhouette examples from a camera ring but consider much fewer training views and instead propose to exploit the projective geometry of the scenes to improve the analysis of the input silhouette.

A few previous efforts have been made to exploit this projective geometry. In [37], a method was proposed for view invariant gait recognition: considering a person walking along a straight line (making a constant angle with the image plane), a side-view was synthesized using a homography. In [38], we proposed an algorithm that projected both shape model and input image in a canonical vertical view, orthogonal or parallel to the direction of motion. In the same spirit, a homographic transformation was later employed in [39] to improve human detection in images

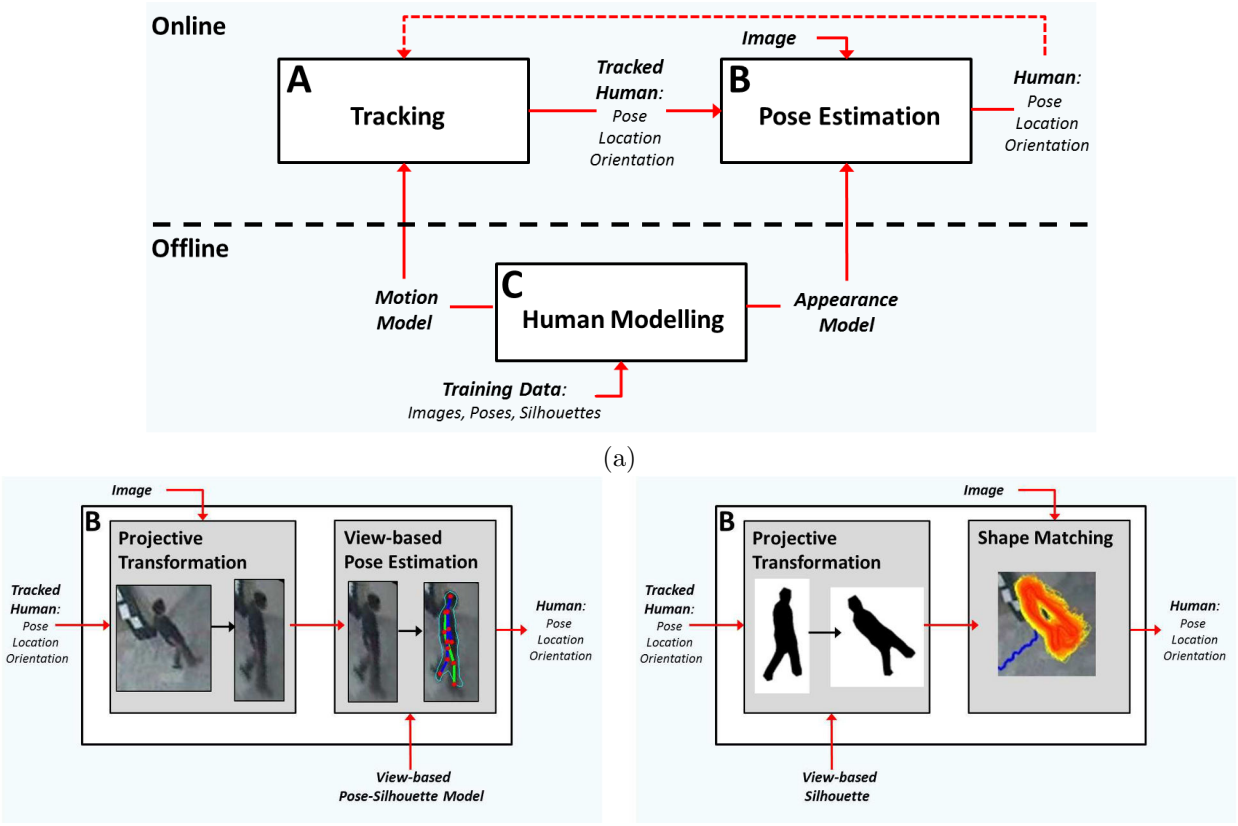


Figure 2: In (a), we show a typical Human Motion Analysis Flowchart made of three different blocks: Tracking (A), Pose Estimation (B) and Human Modelling (C). Tracking and Pose Estimation are performed online while Modelling is offline. In this paper, we focus on the Pose Estimation block where we propose to add a homography-based correction to deal with possible differences in camera viewing angle. This projective transformation can be used in a bottom-up framework (b) to align the input image to the closest training plane and process it with the corresponding view-based model, or in a top-down framework (c) to project a candidate silhouette and match it in the image.

presenting perspective distortion. They reported an improvement in detection rate from 38.3% to 87.2% using 3D scene information instead of scanning over 2D (plus in-plane rotation) on the CAVIAR dataset [21]. The main difference between our proposed approach and these previous methods is that all three only consider the projection to a simple canonical vertical view while we propose to select the closest training view from a given input image and compute the projective transformation between them. Another difference is that we tackle the more complicated task of articulated human pose estimation and propose an extensive numerical evaluation on challenging sequences.

Recent approaches to articulated tracking increasingly more often rely on detailed 3D body models [40, 41]. For these methods, novel camera views do not pose particular problems, as they can generate body appearance for any view through rendering a 3D model. Their main drawback is the computational cost of fitting the thousands of triangles of the mesh models, which makes them less suitable for practical real-time applications such as surveillance. Other existing motion analysis systems [5, 7, 15] usually assume that the camera axis is parallel to the ground and that the observed people are vertically displayed, i.e., elevation angle $\varphi = 0$ and rotation angle $\gamma = 0$ (see Fig. 1b for angle definition), thus making the problem substantially easier. Many authors proposed to discretize the camera viewpoint in a circle around the subjects, selecting a set of training values for the azimuth θ: 36 orientations in [7], 16 in [42, 43], 12 in [5] and 8 in [**?** 15, 44, 45]. Although qualitative results have been presented for street views [7, 15], numerical evaluation is usually conducted using standard testing datasets in laboratory type environments (e.g., HumanEva [46]), and training and testing images are generally captured with a similar camera tilt angle. Very few tackle the problem of pose tracking in surveillance scenarios with low resolution and high perspective distortion as we do. Most of these state-of-the-art systems follow a common basic flowchart (Fig. 2a) with an offline learning stage which consists in building a model of the human appearance from a set of training views, and an online process where the humans are tracked and their pose estimated. These methods explicitly [13, 47] or implicitly [7, 15] apply a similarity transformation between their models and the processed images, most of the time with only scale and translation elements (no in-plane rotation). Few pose tracking algorithms exploit the key constraints provided by scene calibration. A part from our previous work [48], the only example of tracking results in

crowded video-surveillance sequences with perpective effect was presented in [49] but no body pose was estimated.

Even though many new types of image features have recently been developed, silhouette-based approaches are still receiving much attention. These approaches focus on the use of the binary silhouette of the human body as a feature for detection [20, 50], tracking [47, 51, 52, 53, 54], pose estimation [2, 5, 6, 7, 45, 55] or action recognition [56, 57] to cite a few. They rely on the observation that most human gestures can be recognized using only the outline shape of the body silhouette. The most important advantage of these features is their ease of extraction from raw video frames using low-level processing tasks like background subtraction or edge detection algorithms. However, in presence of perspective effect, the distortion will cause the parts of the subject that are closer to the lens to appear abnormally large, thus deforming the shape of the human contour in ways that can prevent a correct analysis as discussed in [58]. In this paper, we show how projective geometry can be exploited to improve silhouette-based approaches in such cases.

1.2. Overview of the Approach.

We present a methodology for view-invariant monocular human motion analysis in man-made environments in which we assume that observed people move on a known ground plane, a valid assumption in most surveillance scenarios. Considering the framework depicted in Fig. 2a, in this paper we focus on the pose estimation block (B). In our previous work, we dealt with the other two blocks (modelling and tracking). Interested readers are encouraged to consult [13, 45] for bottom-up methods and [48] for our top-down tracking framework, or [59] for a more complete discussion.

The basic idea of this paper is that projective geometry could be exploited to compensate for the difference of camera view between input and training images. In our earlier work [38], we gave a first insight on how the use of a projective transformation for shape registration improves silhouette-based pose estimation. In this paper, we consider a wider range of possible viewpoints and propose to estimate the homography transformation between training and test views. This homography is used to compensate for changes in silhouette due to the novel unseen viewpoint. First, the position of the camera with respect to the observed object, the view, is parameterized with two angles, *latitude* φ and *longitude* θ, that define the upper viewing hemisphere as shown in Fig. 1b. Our proposal then relies on two separate stages: The **off-line stage** consists in discretizing the viewing hemisphere into a reduced number of training viewpoints. In this paper, we use the MoBo dataset for training and consider 8 training viewpoints uniformly distributed around the subject. For each view, body poses and silhouettes are labelled (Fig. 3) and used to train a model/mapping between view-based silhouette shape and pose. Camera and scene calibrations are also performed in training and testing views.

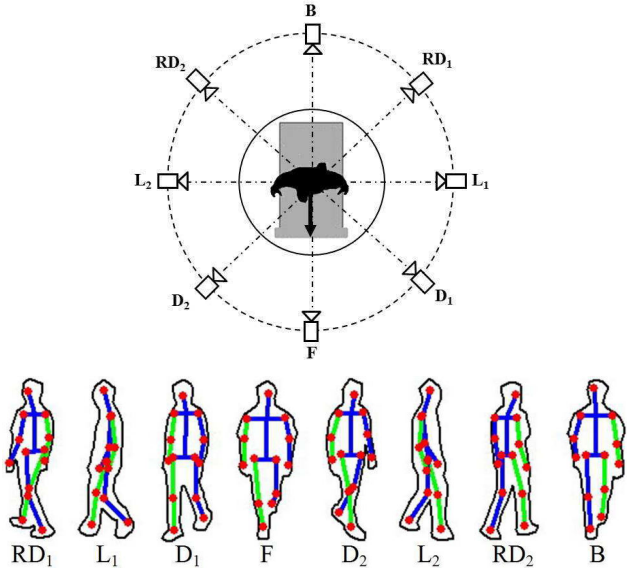


Figure 3: (up) Viewpoint discretization: in this work, we use the MoBo dataset [22] and discretize the viewing hemisphere into 8 locations where θ is uniformly distributed around the subject. Examples of training images are given in Fig. 1c for lateral (L1), diagonal (D1), front (F), rear-diagonal (RD2) and back (B) views. (down) 8 view-based shapes and 2D poses of a particular training snapshot.

Given a test image, the **online stage** consists in selecting the closest training view on the viewing hemisphere and computing the homography that relates the corresponding training plane to the image points using the dominant 3D directions of the scene, the location on the ground plane and the camera view in both training and input images. This transformation can potentially compensate for the effect of both discretization along θ and variations along φ, and removes part of the perspective effect. This homographic transformation can be used 1) in a *bottom-up* manner to align an input image to a selected training plane for a view-based processing (see Fig. 2b) or 2) *top-down* to project a candidate silhouette and match it directly in the image (see Fig. 2c). The work presented in this paper can be seen as an extension of [60]. We take steps toward detailling and generalizing the algorithm and analyze the improvement obtained when the proposed projective transformation is included within two different type of motion analysis frameworks [45, 48]. We have also carried out an exhaustive experimentation to validate our approach with a numerical evaluation and present a comparison with the state-of-the-art approach which consists in using a four-parameter similarity transform. Standard testing data sets for pose estimation (e.g., HumanEva [46]) do not consider perspective distortion and can not be used in this paper to offer a comparison with state-of-the-art work. We instead employ the CAVIAR dataset [21] that presents very challenging sequences with perspective distortion (see Fig. 3a). Our qualitative and quantitative results on this dataset, in both *bottom-up* and *top-down* frameworks, demonstrate the significant improvements of

the proposed homographic alignment for silhouette-based human motion analysis.

The rest of the paper is organized as follows. First, some geometrical considerations are explained in Section 2. The computation of the projective transformation is then described in Section 3 while the qualitative and quantitative evaluations are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. Geometrical Considerations in Man-Made Environments

We propose to exploit camera and scene knowledge when working in a man-made environments which is the case of most video-surveillance scenarios.

2.1. Notations

In the following sections upper case letters, e.g., $\mathbf{X}$ or X, will be used to indicate quantities in space whereas image quantities will be indicated with lower case letters, e.g., $\mathbf{x}$ or x. Euclidean vectors are denoted with upright boldface letters, e.g $\mathbf{x}$ or $\mathbf{X}$, while slanted letters denote their cartesian coordinates, e.g., (x, y, z) or (X, Y).

Following notations used in [61], underlined fonts $\underline{\bullet}$ indicate homogeneous coordinates in projective spaces, e.g $\underline{\mathbf{x}}$ or $\underline{\mathbf{X}}$. A homogeneous point $\underline{\mathbf{x}} \in \mathbb{P}^n$ is composed of a vector $\mathbf{m} \in \mathbb{R}^n$ and a scalar ρ (usually referred to as the homogeneous part):

$$\underline{\mathbf{x}} = \begin{bmatrix} \mathbf{m} \\ \rho \end{bmatrix} \in \mathbb{P}^n \subset \mathbb{R}^{n+1}, \tag{1}$$

where the choice $\rho = 1$ is the original Euclidean point representation while $\rho = 0$ defines the points at infinity. The homogeneous point $\underline{\mathbf{x}}$ thus refers to the Euclidean point $\mathbf{x} \in \mathbb{R}^n$:

$$\mathbf{x} = \mathbf{m}/\rho. \tag{2}$$

By definition, all the homogeneous points $\{[\rho\mathbf{x}^T, \rho]^T\}_{\rho \in \mathbb{R}^*}$ represent the same Euclidean point $\mathbf{x}$ (see [62]) and, for homogeneous coordinates, "=" means an assignment or an equivalence up to a non-zero scale factor.

2.2. Camera and Scene Calibration

Supposing observed humans are walking on a planar ground floor with a vertical posture, camera model and ground plane assumptions provide useful geometric constraints that help reducing the search space as in [20, 49, 63], instead of searching for all scales, all orientations and all positions. During the scene calibration two 3×3 homography matrices are calculated: $\mathbf{H}_\mathrm{g}$ which characterizes the mapping between the ground plane in the image and the real world ground plane Π_gd and $\mathbf{H}_\mathrm{h}$ relating the head plane in the image with Π_gd. In this work, the homography matrices are estimated by the least-squares method using four or more pairs of manually preannotated points in several frames. The 2 homography mappings are illustrated in Fig. 4. Note that when surveillance cameras with a high field of view are used (as with [21]), a previous lens calibration is required to correct the optical distortion.

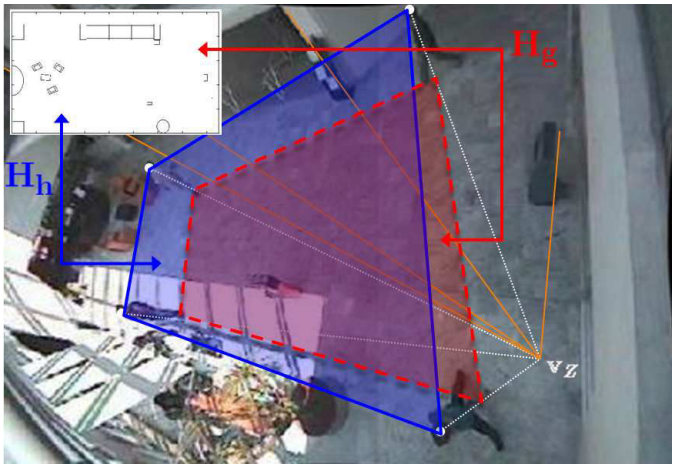


Figure 4: Camera and Scene Calibration: 2 homography matrices are calculated from manual annotations: $\mathbf{H_g}$ characterizing the mapping between the ground plane in the image (*red dashed line*) and the real world ground plane Π_gd (*upper left*) and $\mathbf{H_h}$ relating the head plane in the image (*blue solid line*) with Π_gd. The vertical vanishing point $\mathbf{v}_Z$ and the horizontal vanishing line are also computed using the straight lines from walls and floor observed in the scene.

Given an estimate of the subject's location (X, Y) on the world ground plane Π_gd, the planar homographies $\mathbf{H}_\mathrm{g}$ and $\mathbf{H}_\mathrm{h}$ are used to evaluate the location of the subject's head $\mathbf{x}_H$ and "feet" $\mathbf{x}_F$ in the image I:

$$\underline{\mathbf{x}}_H = \mathbf{H}_\mathrm{h} \cdot [X, Y, 1]^T, \tag{3}$$

$$\underline{\mathbf{x}}_F = \mathbf{H}_\mathrm{g} \cdot [X, Y, 1]^T, \tag{4}$$

where points in the projective space $\mathbb{P}^2$ are expressed in homogeneous coordinates.

In this work, we want to compensate for the difference of camera view between input and training images using the dominant 3D directions of the scenes. We suppose that the camera model is known and people walk in a structured man-made environment where straight lines and planar walls are plentiful. The transformation matrices introduced in the next section are calculated online using the vanishing points [1] evaluated in an off-line stage: the positions of the vertical vanishing point $\mathbf{v}_Z$ and $\mathbf{l}$, the vanishing line of the ground plane, are directly obtained after a manual annotation of the parallel lines (on the ground and walls) in the image. An example of vertical vanishing point localization is given in Fig. 4. This method makes sense only for man-made environments because of the presence of numerous easy-to-detect straight

[1] A vanishing point is the intersection of the projections in the image of a set of parallel world lines. Any set of parallel lines on a plane define a vanishing point and the union of all these vanishing points is the vanishing line of that plane [64].

lines. Previous work for vanishing points detection [65] could be used to automate the process.

Once we have calibrated the camera in the scene, the camera cannot be moved, which is a limitation of the proposal. In practice, the orientation of the camera could change, for example, due to the lack of stability of the camera support. A little change in orientation has a great influence in the image coordinates, and therefore invalidates previous calibration. However, if the camera is not changed in position, or position change is small with respect to the depth of the observed scene, the homography can easily be re-calibrated automatically. An automatic method to compute homographies and line matching between image pairs like the one presented in [66] can then be used.

3. Projective Transformation for View-invariance

As demonstrated in [37], for objects far enough from the camera, we can approximate the actual 3D object as being represented by a planar object. In other words, a person can be approximated by a planar object if he or she is far enough from the camera [2]. As shown in [58], in the presence of perspective distortion neither similarity nor affine model provide reasonable approximation for the transformation between a prior shape and a shape to segment. The authors demonstrated that a planar projective transformation is a better approximation even though the object shape contour is roughly planar (e.g., a toy elephant). Following these two observations, we propose to find a projective transformation, i.e., a homography, between training and testing camera views to compensate for the effect of both discretization along θ and variations along φ, thus alleviating the effect of perspective distortion on silhouette-based human motion analysis.

3.1. Projection to Vertical Plane

Following the classical notation of 3D projective geometry [62], a 3D point (X, Y, Z) is related to its 2D image projection $\mathbf{x}$ via a 3×4 projection matrix $\mathbf{M}$:

$$\underline{\mathbf{x}} = \mathbf{M} \cdot [X, Y, Z, 1]^T, \tag{5}$$

where $\underline{\mathbf{x}} \in \mathbb{P}^2$. The projective transformation matrix $\mathbf{M}$ can be determined with a series of intrinsic and extrinsic parameters or, as shown in [64], it can be defined as a function of the vanishing points of the dominant 3D directions.

Suppose we want to relate the image I with a vertical plane Π ($\Pi \perp \Pi_{\text{gd}}$), whose intersection with the ground plane Π_{gd} is $\mathbf{G}$. The plane Π is thus spanned by the vertical Z-axis and horizontal G-axis. In that sense, (5) becomes:

$$\underline{\mathbf{x}} = \mathbf{H}_{I \leftarrow \Pi} \cdot [G, Z, 1]^T, \tag{6}$$

[2] This hypothesis is obviously not strictly true as it does not depend solely on the distance to the camera but also on the pose and orientation of the person w.r.t. the camera

with G a coordinate on the G-axis and $\mathbf{H}_{I \leftarrow \Pi}$ a 3×3 homography matrix that can be computed from the vanishing points of the dominant 3D directions of Π :

$$\mathbf{H}_{I \leftarrow \Pi} = [\mathbf{v}_G \;\; \alpha \mathbf{v}_Z \;\; \mathbf{o}]. \tag{7}$$

where $\mathbf{v}_Z$ is the vertical vanishing point, $\mathbf{o}$ is the origin of the world coordinate system and α is a scale factor. $\mathbf{v}_G$ is the horizontal vanishing point of plane Π in I i.e., the vanishing point along the horizontal direction $\mathbf{G}$ in image I. This vanishing point $\mathbf{v}_G$ can be localized as the intersection of line $\mathbf{g}$, the projection of $\mathbf{G}$ in the image I and $\mathbf{l}$, the horizontal vanishing line in I:

$$\mathbf{v}_G = \mathbf{l} \times \mathbf{g}, \tag{8}$$

where $\times$ represents the vector product, and $\mathbf{l}$ is the vanishing line of the ground plane (see [62] for details). Two examples of horizontal vanishing point localizations are given in Fig. 5.

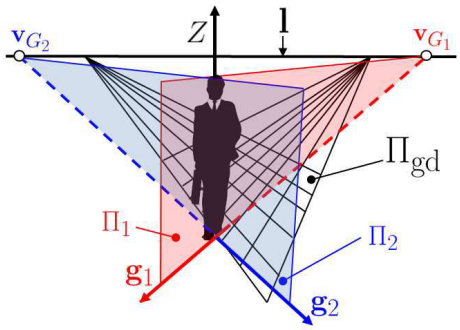


Figure 5: Horizontal vanishing point localization for homography to vertical plane centered on the human body: 2 examples are given for 2 different directions $\mathbf{g}_1$ and $\mathbf{g}_2$ on the ground plane Π_{gd}. Π_1 is the vertical plane parallel to the real-world direction G_1 and Π_2 the one parallel to G_2. The vanishing points $\mathbf{v}_{G_1}$ and $\mathbf{v}_{G_2}$ are the intersection points of $\mathbf{g}_1$ and $\mathbf{g}_2$ with the horizon line $\mathbf{l}$, i.e., the vanishing line of the ground plane.

3.2. Projection Image-Training View Through a Vertical Plane

The 3×3 transformation $\mathbf{P}_{I_2 \Pi I_1}$ between two images I_1 and I_2 through a vertical plane Π observed in both images can be obtained as the product of 2 homographies defined up to a rotational ambiguity. The first one, $\mathbf{H}_{\Pi \leftarrow I_1}$, projects the 2D image points in I_1 to the vertical plane Π and the other one, $\mathbf{H}_{I_2 \leftarrow \Pi}$, relates this vertical plane to the image I_2. We thus obtain the following equation that relates the points $\mathbf{x}_1$ from I_1 with image points $\mathbf{x}_2$ from I_2:

$$\underline{\mathbf{x}}_2 = \mathbf{P}_{I_2 \Pi I_1} \cdot \underline{\mathbf{x}}_1, \tag{9}$$

where $\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2 \in \mathbb{P}^2$ and with:

$$\begin{aligned} \mathbf{P}_{I_2 \Pi \; I_1} &= \mathbf{H}_{I_2 \leftarrow \Pi} \cdot \mathbf{H}_{\Pi \leftarrow I_1} \\ &= \mathbf{H}_{I_2 \leftarrow \Pi} \cdot (\mathbf{H}_{I_1 \leftarrow \Pi})^{-1}. \end{aligned} \tag{10}$$

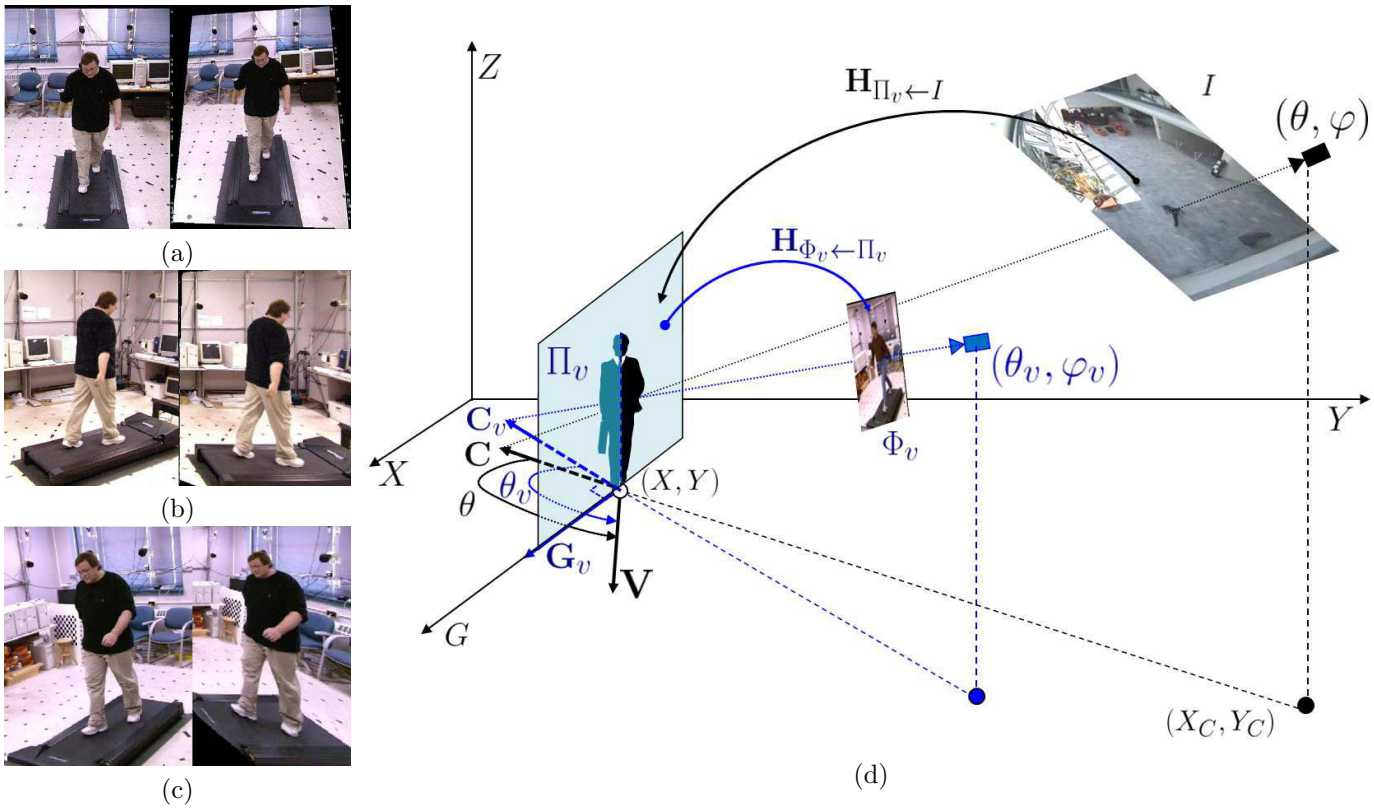


Figure 6: (a) Projection on the vertical plane: examples of original and warped images resulting from applying the homography $\mathbf{H}_{\Pi_v \leftarrow \Phi_v}$ for frontal (a), "rear-diagonal" (b) and diagonal (c) views of the MoBo dataset. (d) Schematical representation of the transformation between 2 images through a vertical plane: testing image I and training image plane Φ_v can be related through a vertical plane Π_v. The transformation $\mathbf{P}_{\Phi_v \Pi_v\, I}$ is obtained as the product of $\mathbf{H}_{\Phi_v \leftarrow \Pi_v}$ and $\mathbf{H}_{\Pi_v \leftarrow I}$ while the inverse projection $\mathbf{P}_{I \Pi_v \Phi_v}$ can be obtained as the product of $\mathbf{H}_{I \leftarrow \Pi_v} = (\mathbf{H}_{\Pi_v \leftarrow I})^{-1}$ and $\mathbf{H}_{\Pi_v \leftarrow \Phi_v} = (\mathbf{H}_{\Phi_v \leftarrow \Pi_v})^{-1}$.

Following Eq. 7, the two homographies $\mathbf{H}_{I_1 \leftarrow \Pi}$ and $\mathbf{H}_{I_2 \leftarrow \Pi}$ can be computed from the vanishing points of the 3D directions spanning the vertical plane Π i.e., the vertical Z-axis and the reference horizontal line $\mathbf{G} = \Pi \wedge \Pi_{\text{gd}}$, intersection of Π and ground plane Π_{gd}.

In the same way, we now want to relate 2 images, e.g., training and testing images, observing two different calibrated scenes with 2 different subjects performing the same action from two similar viewing angles. These images can potentially be related through a vertical plane centered in the human body following Eq. 9 . The problem is to select the vertical plane that will optimize the 2D shape correspondence between the 2 images. We choose to select this vertical plane in the training image, where the azimuth angle θ is known and the camera is in an approximately horizontal position (i.e., elevation angle $\varphi \approx 0$), and consider the closest vertical plane centered on the human body: if a camera view Φ is defined by its azimuth and elevation angles (θ, φ) on the viewing hemisphere (Fig. 3a), the closest vertical plane Π is the plane defined as $(\theta, 0)$.

Thus, considering a set of training views $\{\Phi_v\}_{v=1}^{N_v}$, the associated homographies $\{\mathbf{H}_{\Phi_v \leftarrow \Pi_v}\}_{v=1}^{N_v}$ relating each view and its closest vertical plane Π_v centered on the human body are computed during the off-line stage (following Eq. 7) and stored for online use[3]. Each vertical plane Π_v is spanned by the vertical Z-axis and a reference horizontal vector $\mathbf{G}_v \in (\Pi_v \wedge \Pi_{\text{gd}})$. Examples of projection on a vertical plane are given for 3 of the 8 MoBo training views in Fig. 6. The perspective distortion, particularly severe in the front view (large head and short legs), is corrected: the image appears distorted but the global figure recovers real morphological proportions in the front view (Fig. 6a) while we can observe how the transformation tends to place the feet at the same vertical position and remove the perspective effect for the rear-diagonal (Fig. 6b) view.

Given a test image I with an observed human at location (X, Y) on the ground plane Π_{gd}, the azimuth $\theta \in [-\pi, \pi]$ (i.e., camera viewpoint or the subject's orientation w.r.t. the camera) is defined on the ground as:

$$\theta = \widehat{\mathbf{CV}}, \tag{11}$$

where vectors $\mathbf{C}$ and $\mathbf{V} \in \mathbb{R}^2$ are the projections on the ground plane Π_{gd} of the camera viewing direction and the orientation vector respectively[4]. The viewing direction

[3]The training views considered in this work are not exactly frontal explaining why $\mathbf{H}_{\Pi_v \leftarrow \Phi_v}$ are taken into account.

[4]The angle θ is π when the subject is facing the camera and θ is 0 when facing away.

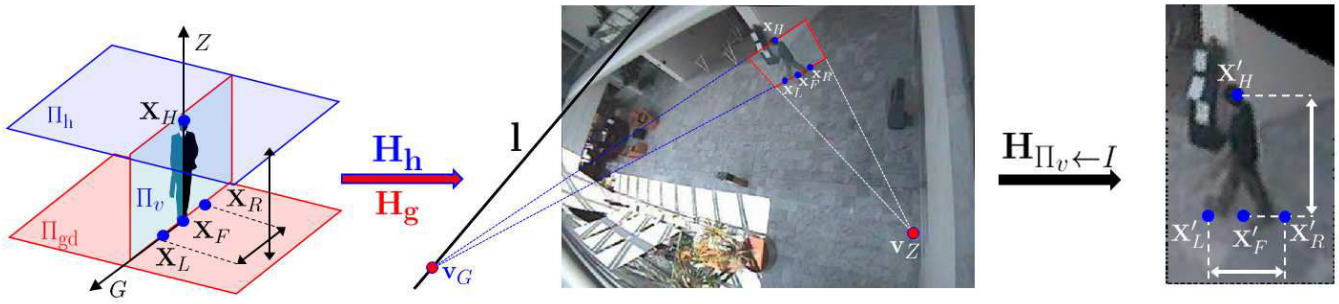


Figure 7: Projection to Vertical Plane. Four real world coplanar points are selected on Π_v : $\mathbf{X}_L$, $\mathbf{X}_R$, $\mathbf{X}_F$ on the ground plane Π_{gd} along the G-axis and $\mathbf{X}_H$ the center of the head (*left*). The four points are reprojected in the image I obtaining $\underline{\mathbf{x}}_L = \mathbf{H}_g\underline{\mathbf{X}}_L$, $\underline{\mathbf{x}}_R = \mathbf{H}_g\underline{\mathbf{X}}_R$, $\underline{\mathbf{x}}_F = \mathbf{H}_g\underline{\mathbf{X}}_F$ and $\underline{\mathbf{x}}_H = \mathbf{H}_h\underline{\mathbf{X}}_H$ (*middle*). The image points along the G-axis are then used to localize the vanishing point $\mathbf{v}_G$. The homography $\mathbf{H}_{\Pi_v \leftarrow I}$ relating the input image with the selected vertical plane is then obtained from vanishing points $\mathbf{v}_Z$ and $\mathbf{v}_G$ following Eq. 7. The scale factor α in $\mathbf{H}_{\Pi_v \leftarrow I}$ is then computed so that the height to width ratio stays constant between the set of reprojected points $\{\mathbf{X}'_L, \mathbf{X}'_R, \mathbf{X}'_F, \mathbf{X}'_H\}$ (*right*) and the original real-world points $\{\mathbf{X}_L, \mathbf{X}_R, \mathbf{X}_F, \mathbf{X}_H\}$ (*left*).

is defined as the line connecting the subject and camera (originating from the camera center) and the orientation direction is a vector perpendicular to the shoulder line of the subject pointing in the direction he or she is facing (see Fig. 6). Note that $\mathbf{C}$ is easily evaluated as:

$$\mathbf{C} = \begin{bmatrix} X - X_C \\ Y - Y_C \end{bmatrix}, \tag{12}$$

where (X_C, Y_C) is the projection on the ground plane of the camera center [5]. The direction of $\mathbf{V}$ can be found by rotating $\mathbf{C}$ around the Z-axis if θ is known:

$$\mathbf{V} \propto \mathbf{R}(\theta) \cdot \mathbf{C}, \tag{13}$$

where $\mathbf{R}(.)$ denotes a 2×2 rotation matrix.

Table 1: Azimuth $\theta_v = \widehat{\mathbf{C}_v\mathbf{V}}$ and $\widehat{\mathbf{VG}_v}$ angle defining the vertical plane Π_v for the 8 training viewpoints of the MoBo dataset (Fig. 3): lateral (L_1 & L_2), diagonal (D_1 & D_2), rear-diagonal (RD_1 & RD_2), front (F) and back (B) views.

	View							
	RD_1	L_1	D_1	F	D_2	L_2	RD_2	B
θ_v	$\frac{\pi}{4}$	$\frac{\pi}{2}$	$\frac{3\pi}{4}$	π	$-\frac{3\pi}{4}$	$-\frac{\pi}{2}$	$-\frac{\pi}{4}$	0
$\widehat{\mathbf{VG}_v}$	$\frac{\pi}{4}$	0	$-\frac{\pi}{4}$	$-\frac{\pi}{2}$	$-\frac{3\pi}{4}$	π	$\frac{3\pi}{4}$	$\frac{\pi}{2}$

Given $\{\theta_v = \widehat{\mathbf{C}_v\mathbf{V}}\}_{v=1}^{N_v}$ the N_v training values for θ (c.f. Tab. 1) and given an estimation of θ for the observed subject, a training view Φ_v is selected so that:

$$v = \arg \min_{v \in \{1, N_v\}} |\theta - \theta_v|. \tag{14}$$

The transformation $\mathbf{P}_{\Phi_v \Pi_v\ I}$ (illustrated in Fig. 6) between input image I and Φ_v through the vertical plane Π_v can then potentially be obtained as the product:

$$\mathbf{P}_{\Phi_v \Pi_v\ I} = \mathbf{H}_{\Phi_v \leftarrow \Pi_v} \cdot \mathbf{H}_{\Pi_v \leftarrow\ I}, \tag{15}$$

[5] As indicated in [62], the vanishing point is the image of the vertical "footprint" of the camera centre on the ground plane, i.e., $\underline{\mathbf{X}}_C = (\mathbf{H}_g)^{-1} \cdot \underline{\mathbf{v}}_Z$ with $\mathbf{X}_C = (X_C, Y_C)$.

up to a rotational ambiguity. The problem now consists of finding the plane Π_v in the image I, i.e., the vanishing points of the 3D directions, and compute $\mathbf{H}_{\Pi_v \leftarrow\ I} = (\mathbf{H}_{\ I \leftarrow \Pi_v})^{-1}$ from Eq. 7. The plane Π_v is spanned by the vertical Z-axis and a horizontal axis $\mathbf{G} = \mathbf{G}_v$ which can be found in the real 3D world by rotating $\mathbf{V}$ about the Z-axis:

$$\mathbf{G} \propto \mathbf{R}(\widehat{\mathbf{VG}_v}) \cdot \mathbf{V}. \tag{16}$$

The training values for $\widehat{\mathbf{VG}_v}$ are given in Tab. 1. Two real world 3D points $\mathbf{X}_L$, $\mathbf{X}_R$ are then selected on the ground floor along the G-axis at each side of the subject (see Fig. 7a). In practice, we select 2 points at 50 cm from the subject. $\mathbf{X}_L$ and $\mathbf{X}_R$ are then reprojected in the image I obtaining $\underline{\mathbf{x}}_L = \mathbf{H}_g\underline{\mathbf{X}}_L$ and $\underline{\mathbf{x}}_R = \mathbf{H}_g\underline{\mathbf{X}}_R$, where $\underline{\mathbf{X}}_L, \underline{\mathbf{X}}_R \in \mathbb{P}^2$ are expressed in the ground plane coordinates. These two image points can be used to localize the vanishing point $\mathbf{v}_G$ along real-world G-axis in the image (Fig. 7b) as follows:

$$\underline{\mathbf{v}}_G = (\underline{\mathbf{x}}_L \times \underline{\mathbf{x}}_R) \times \mathbf{l}, \tag{17}$$

where $\times$ represents the vector product, and $\mathbf{l} \in \mathbb{P}^2$ is the vanishing line of the ground plane (see [62] for details).

The computation of $\mathbf{H}_{\Pi_v \leftarrow\ I}$ relating the input image with the selected vertical plane is then obtained following Eq. 7. The scale factor α in Eq. 7 is evaluated using four known coplanar points[6] in the real-world vertical plane Π_v: $\mathbf{X}_L$, $\mathbf{X}_R$ (from above), the subject's ground floor location $\mathbf{X}_F$ and $\mathbf{X}_H$, the center of the subject's head, i.e., the vertical projection on the head plane Π_h of the ground floor location (see Fig. 7a). The images $\underline{\mathbf{x}}_L = \mathbf{H}_g\underline{\mathbf{X}}_L$, $\underline{\mathbf{x}}_R = \mathbf{H}_g\underline{\mathbf{X}}_R$, $\underline{\mathbf{x}}_F = \mathbf{H}_g\underline{\mathbf{X}}_F$ and $\underline{\mathbf{x}}_H = \mathbf{H}_h\underline{\mathbf{X}}_H$ of these four points in I (Fig. 7b) are reprojected in the plane Π_v using $\mathbf{H}_{\Pi_v \leftarrow\ I}$ obtaining $\mathbf{X}'_L$, $\mathbf{X}'_R$, $\mathbf{X}'_F$ and $\mathbf{X}'_H \in \mathbb{R}^2$ (Fig. 7c). The scale factor α in $\mathbf{H}_{\Pi_v \leftarrow\ I}$ is then computed so that the height to width ratio stays constant between the set

[6] Note that even if four points have been considered in our implementation, three points would be sufficient.

of reprojected points $\{\mathbf{X}'_L, \mathbf{X}'_R, \mathbf{X}'_F, \mathbf{X}'_H\}$ and the original real-world points $\{\mathbf{X}_L, \mathbf{X}_R, \mathbf{X}_F, \mathbf{X}_H\}$, i.e.,

$$\frac{||\mathbf{X}_H - \mathbf{X}_F||}{||\mathbf{X}_R - \mathbf{X}_L||} = \frac{||\mathbf{X}'_H - \mathbf{X}'_F||}{||\mathbf{X}'_R - \mathbf{X}'_L||} = \frac{||h_\alpha(\mathbf{x}'_H) - h_\alpha(\mathbf{x}'_F)||}{||h_\alpha(\mathbf{x}'_R) - h_\alpha(\mathbf{x}'_L)||}, \tag{18}$$

where, for ease of notation, we define the one-to-one mapping function $h_\alpha : \mathbb{R}^2 \mapsto \mathbb{R}^2$ which transforms image points to plane Π_v using the homography $\mathbf{H}_{\Pi_v \leftarrow\ I}$: $\mathbf{X} = h_\alpha(\mathbf{x}) \Leftrightarrow \underline{\mathbf{X}} = \mathbf{H}_{\Pi_v \leftarrow\ I} \cdot \underline{\mathbf{x}}$. In our case, we assume the head is at 170 cm from the floor (average human height) and select $\mathbf{X}_L$ and $\mathbf{X}_R$ to be 100 cm apart, we thus have to find α which minimizes:

$$E(\alpha) \triangleq \left| 1.7 - \frac{||h_\alpha(\mathbf{x}'_H) - h_\alpha(\mathbf{x}'_F)||}{||h_\alpha(\mathbf{x}'_R) - h_\alpha(\mathbf{x}'_L)||} \right|, \tag{19}$$

i.e., a convex optimization problem which is easily solved by gradient descent search.

Finally, once $\mathbf{H}_{\Pi_v \leftarrow\ I}$ has been calculated, $\mathbf{P}_{\Phi_v \Pi_v\ I}$ can be computed using Eq. 15. The rotational ambiguity in choosing the coordinate system is resolved using the same four points and checking that the vectors $\mathbf{U}, \mathbf{U}' \in \mathbb{R}^3$ resulting from the two cross products $\mathbf{U} = \langle \mathbf{X}_L \mathbf{X}_R \rangle \times \langle \mathbf{X}_F \mathbf{X}_H \rangle$ and $\mathbf{U}' = \langle \mathbf{X}'_L \mathbf{X}'_R \rangle \times \langle \mathbf{X}'_F \mathbf{X}'_H \rangle$ point in the same direction, otherwise the G-axis is flipped in matrix $\mathbf{H}_{\Pi_v \leftarrow\ I}$. Eq. 15 becomes:

$$\mathbf{P}_{\Phi_v \Pi_v\ I} = \begin{cases} \mathbf{H}_{\Phi_v \leftarrow \Pi_v} \cdot \mathbf{H}_{\Pi_v \leftarrow\ I} & \text{if } \mathbf{U} \cdot \mathbf{U}' \geq 0, \\ \mathbf{H}_{\Phi_v \leftarrow \Pi_v} \cdot \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \mathbf{H}_{\Pi_v \leftarrow I} & \text{otherwise.} \end{cases} \tag{20}$$

The entire process leading to the computation of the projective transformation $\mathbf{P}_{\Phi_v \Pi_v\ I}$ is summarized in Alg. 1.

Algorithm 1: Projection Image to Training View.

input : Triplet (X, Y, θ).
output: Projective Transformation $\mathbf{P}_{\Phi_v \Pi_v\ I}$.

- Select the training view Φ_v (Eq. 14);
- Compute camera viewing direction $\mathbf{C}$ (Eq. 12) ;
- Find orientation vector $\mathbf{V}$ (Eq. 13);
- Find the real-world G-axis defining Π_v (Eq. 16);
- Localize the vanishing point $\mathbf{v}_{\ G}$ using Eq. 17;
- Calculate $\mathbf{H}_{\Pi_v \leftarrow\ I} = (\mathbf{H}_{\ I \leftarrow \Pi_v})^{-1}$ using Eq. 7;
- Compute the scale factor α (Eq. 19);
- Calculate $\mathbf{P}_{\Phi_v \Pi_v\ I}$ using Eq. 20;

4. Experiments

We now experimentally validate the proposed projective transformation within two different types of motion analysis framework. In section 4.1, we consider a bottom-up image analysis scheme where the ground plane position (X, Y) and the camera viewpoint θ are estimated deterministically using a Kalman filter. Our transformation $\mathbf{P}_{\Phi_v \Pi_v\ I}$ is then used to align the input image to one of the training planes and process it with the corresponding view-based silhouette model. In section 4.2, we employ our homography based alignment within a top-down pose tracking framework where X, Y, θ and a body pose parameter are sampled and estimated stochastically. The corresponding candidate silhouettes are then transformed using the inverse projection $\mathbf{P}_{\ I \Pi_v \Phi_v} = (\mathbf{P}_{\Phi_v \Pi_v\ I})^{-1}$ and matched in the original input image. In both cases, we use the 8 training viewpoints from the Mobo dataset for training and the annotated sequences from Caviar [21] for testing.

4.1. Image Transformation and Bottom-up Analysis

One of the bottlenecks of deterministic frameworks is that the estimation of the location can be relatively noisy. First, in Sect. 4.1.1, we analyze how well our projective transformation can potentially work in case of a perfect tracker that we simulate using ground truth data. We conduct a qualitative evaluation employing manually labelled head locations to generate ground truth data for triplets (X, Y, θ) in several sequences. Next, we apply our homography within a real bottom-up framework from [45] in Sect. 4.1.2. Finally, we numerically evaluate how well our homographic alignment can work with a noisy tracker in Sect. 4.1.3 and discuss the different results in Sect. 4.1.4.

4.1.1. Qualitative results using ground-truth data

A series of gait sequences are first selected from Caviar: in these sequences people are walking in various directions and the changing perspective effect can be observed. For each sequence, the trajectory $\{X_t, Y_t\}_{t=1}^{N_t}$ on the ground floor is directly recovered from the manual labelling using $\mathbf{H}_\mathrm{h}$ which relates the head plane in the image with the ground plane Π_gd. Supposing that the subject is facing in the direction of motion, we estimate the direction $\mathbf{V}_t$ and consequently the viewpoint angle θ_t at time t from the trajectory $\{X_t, Y_t\}_{t=1}^{N_t}$:

$$\theta_t = \arccos\left(\frac{\mathbf{C}_t \cdot \mathbf{V}_t}{||\mathbf{C}_t|| \cdot ||\mathbf{V}_t||} \right), \tag{21}$$

with $\mathbf{V}_t = [X_t - X_{t-1}, Y_t - Y_{t-1}]^T$ and $\mathbf{C}_t$ from Eq. 12. Projections on training plane obtained using the resulting data $\{(X_t, Y_t, \theta_t)\}_{t=1}^{N_t}$ are given in Fig. 8. For each presented sequence, we show (from top to bottom) the trajectory in the image and its projection on the real-world ground plane $\{X_t, Y_t\}_{t=1}^{N_t}$, the extracted subimages, the viewpoints $\{\theta_t\}_{t=1}^{N_t}$ with corresponding training views and, finally, the

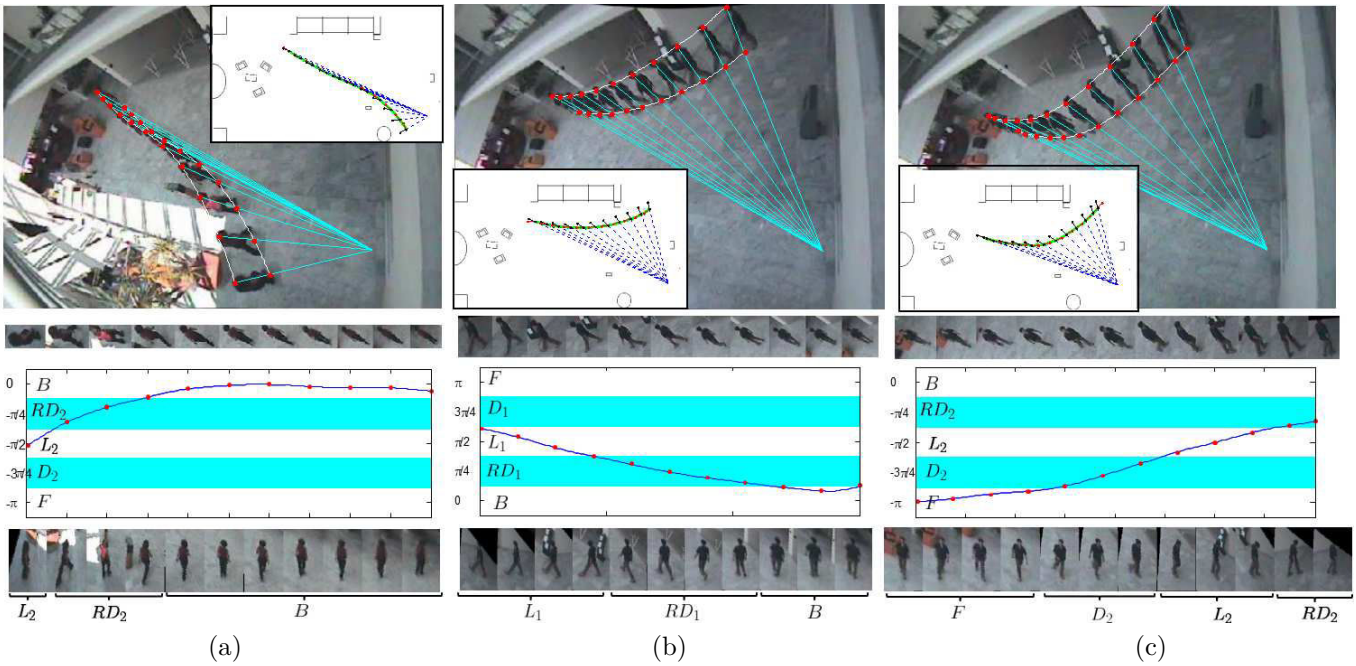


Figure 8: Examples of projections to training planes for *Walk1* (a) and *Walk3* (b and c) sequences [21]. The homographies are computed using "ground truth" locations (X, Y) and viewpoints θ which are estimated from the manual labelling of head location in consecutive frames, the angle θ being estimated from the direction of motion. For each sequence, we show (*from top to bottom*): head and feet trajectories in the image I and corresponding trajectory (X, Y) on the floor with vectors $\mathbf{C}$ and $\mathbf{V}$, the regions of interest, the viewpoint θ and selected training view Φ_v considered to compute $\mathbf{P}_{\Phi_v \Pi_v\ I}$, and finally the warped images $I_{\theta,X,Y}$ for frames $1, 20, \ldots 160, 200$ in (a), frames $1, 15, \ldots, 150$ in (b) and frames $1, 15, \ldots, 150, 160$ in (c).

transformed sub-images $I_{\theta,X,Y}$ for several selected frames. We can observe the smoothness of the different trajectories and how the viewpoint θ slowly changes along the sequences. The regions of interest around the subjects are normalized and projected onto the adequate model plane and the perspective distortion seems corrected. The resulting warped images could be processed using a view-based model.

4.1.2. Bottom-up motion analysis framework

Instead of employing the manual labelling to generate ground truth data for triplets (X, Y, θ) as we previously did in Sec. 4.1.1, we use a head-tracker based on Kalman filter to estimate the head location in consecutive frames and ensure an automatic and reliable estimation of the ground plane trajectory $\{X_t, Y_t\}_{t=1}^{N_t}$. Because of its low shape variability and its top position in the body, the human head is relatively easy to detect, especially in overhead camera views where it is usually less likely to be occluded. Many authors propose computing the vertical histogram of the foreground blob and scanning it, searching for peaks as possible head candidates [26, 63]. This approach is not robust against occlusions and cannot detect the heads "inside" a detection blob as in Fig. 9a. In [67], the authors extend this head candidates search by using a head-shoulder model. Following this approach, we train a similar head shape model and, when given a selected blob (filtered w.r.t its size, position and area), we compute the possible head candidates by searching for local peaks (local maxima) in the direction towards the vertical vanishing point $\mathbf{v}_Z$. We also compute the feet candidates (local minima) and the corresponding probable head location (see Fig. 9b) using the scene calibration, i.e $\underline{\mathbf{x}}_H = \mathbf{H}_{\mathrm{h}} \cdot \mathbf{H}_{\mathrm{g}}{}^{-1} \cdot \underline{\mathbf{x}}_F$. The head shape model is then applied to all the selected head candidates and the confidence weight of each hypothesis is evaluated by edge matching error. Non-human blobs resulting from shadows and reflections are dismissed. An example is given in Fig. 9c.

The system is initialized in the first frames, estimating $\mathbf{x}_H$ by a rough fitting of the silhouette model as in [67]. A tracking is then applied, the state of the each pedestrian, i.e., the ground plane position (X, Y), being estimated at each time step using a Kalman filter. This view-invariant head tracker has shown to be robust, even with difficult cases such as people moving in groups and partial occlusions (see example in Fig. 9d). Again, we suppose that the subject is facing in the direction of motion and estimate the direction $\mathbf{V}_t$ and consequently the viewpoint angle θ_t at time t from the trajectory following Eq. 21. This allows the selection of a training view Φ_v and the computation of the projective transformation $\mathbf{P}_{\Phi_v \Pi_v\ I}$ following Alg. 1. The input image is projected and the resulting warped image $I_{\theta,X,Y}$ is processed to estimate a pose applying the corresponding view-based pose-shape model from our previous work [45]. The system flowchart is presented in Fig. 10.

We processed the gait sequences from Caviar and obtained good results for the sequences where a single in-

Figure 9: Example of a view-invariant head detection with multiple pedestrians from *MeetCrowd* sequence [21]: (a) vertical histogram of the foreground blob, (b) head (crosses) and feet (dots) candidates computed using distance to the vertical vanishing point, (c) detected heads and (d) corresponding trajectories on the ground floor.

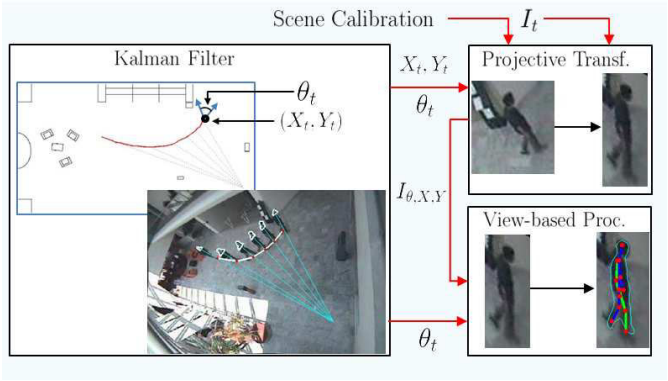


Figure 10: Bottom-up System Flowchart: the state of the each pedestrian, i.e., the ground plane position (X, Y), is estimated at each time step using a Kalman filter whose measure is obtained by projecting vertically the head location found by a view-invariant head detector. At each time step, the camera viewpoint θ_t is estimated using the trajectory of the individual, allowing the selection of a training view. The projective transformation $\mathbf{P}_{\Phi_v \Pi_v\ I}$ relating the corresponding training plane and the image points is calculated using the dominant 3D directions of the scene, the location on the ground plane (X_t, Y_t) and the camera view θ_t. The input image is then projected using this homographic transformation obtaining $I_{\theta,X,Y}$ which is later analysed bottom-up and processed using the corresponding view-based pose-shape models to estimate a silhouette and a pose.

dividual is walking. As expected, the system fails with stationary cases because the viewpoint angle is estimated from the direction of motion. An example of a processed sequence is presented in Fig. 11 where for each presented frame, we show the candidate Region of Interest in the image, the resulting warped image $I_{\theta,X,Y}$ and the obtained pose and silhouette represented ontop of $I_{\theta,X,Y}$. We can observe how the direction of motion slowly changes along the sequence and how the images are projected on the selected model plane. The resulting shapes and poses are reasonably good given the complexity of the task (low resolution and perspective effect). However, the reliability of the warping, and consequently the accuracy of the silhouette and pose estimate, seem to strongly depend on the precision with which both ground plane position (X, Y) and orientation θ are estimated.

4.1.3. Numerical evaluation of the effect of noise

To numerically evaluate this dependence, independently of the possible errors inherent to the tracking algorithm or to the pose estimation technique (e.g., bad initialization or bad model fitting), we conduct a series of simulations using a set of testing ground truth poses $\{\mathbf{k}_1^{GT} \cdots \mathbf{k}_{N_{GT}}^{GT}\}$ and a set of sampled training poses $\{\{\mathbf{k}_i^v\}_{i=1}^{N_T}\}_{v=1}^{N_v}$ (i.e., N_T poses for each training view Φ_v). Each pose is made of 13 hand-labelled 2D joints: $\mathbf{k} = [\mathbf{x}_{k_1}, ..., \mathbf{x}_{k_{13}}] \in \mathbb{R}^{2\times 13}$. For each tested frame $t \in \{1, N_{GT}\}$, we compute the projective transformation $\mathbf{P}_{\ I\,\Pi_v\Phi_v}$ using ground truth location (X_t, Y_t) and viewpoint θ_t from above with additive Gaussian white noises (η_{XY} and η_θ of variance σ_{XY}^2 and σ_θ^2 respectively) and align the N_T training poses $\{\mathbf{k}_1^v \cdots \mathbf{k}_{N_T}^v\}$ from the selected viewpoint Φ_v, obtaining $\{\mathbf{k}_{1,t}^{Hom} \cdots \mathbf{k}_{N_T,t}^{Hom}\}$ with $\forall i \in \{1, N_T\}$:

$$\underline{\mathbf{x}}_{k_j,i,t}^{Hom} = \mathbf{P}_{\ I\,\Pi_v\Phi_v} \cdot \underline{\mathbf{x}}_{k_j,i,t},\ \forall j \in \{1, 13\}. \tag{22}$$

We then compute the average pose error over the testing set taking the closest aligned pose for each frame t:

$$\epsilon^{\text{Hom}} = \frac{1}{N_{GT}} \sum_{t=1}^{N_{GT}} \min_{i\in\{1,N_T\}} d_k(\mathbf{k}_t^{GT}, \mathbf{k}_{i,t}^{Hom}), \tag{23}$$

where d_k, defined as:

$$d_k(\mathbf{k}, \mathbf{k}') \triangleq \frac{1}{13} \sum_{j=1}^{13} ||\mathbf{x}_{k_j} - \mathbf{x}'_{k_j}|| \tag{24}$$

is the average Root Mean Square Error over the 13 2D-joints (called RMS 2D Pose Error from now on).

We repeat the same operation considering a Euclidean 2D similarity transformation $\mathbf{T}$ to align training poses to the tested images as in [47, 13, 5, 7, 15] and compute:

$$\epsilon^{\text{Sim}} = \frac{1}{N_{GT}} \sum_{t=1}^{N_{GT}} \min_{i\in\{1,N_T\}} d_k(\mathbf{k}_t^{GT}, \mathbf{k}_{i,t}^{Sim}), \tag{25}$$

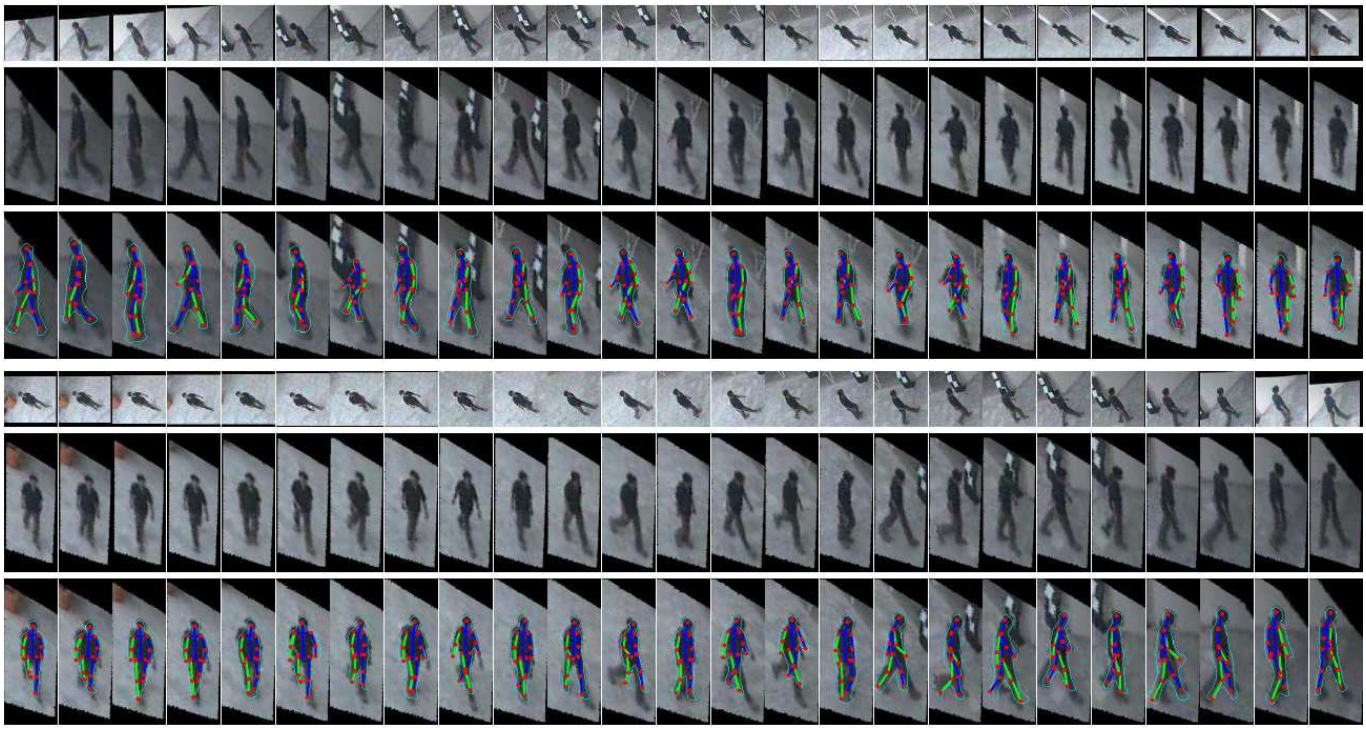


Figure 11: Result obtained using our bottom-up framework on the *Walk3* sequence. For each presented frame, we show the extracted subimage (*top*), the foreground image projected on the selected training plane $I_{\theta,X,Y}$ and the same image with estimated shape and 2D pose after being processed using a view-based model. Results are presented in the attached videos *Walk3b_processed.avi*.

where $\mathbf{k}^{Sim} = [\mathbf{x}_{k_1}^{Sim}, ..., \mathbf{x}_{k_{13}}^{Sim}] \in \mathbb{R}^{2\times 13}$ with:

$$\mathbf{x}_{k_j}^{Sim} = \mathbf{T} \cdot \mathbf{x}_{k_j}, \ \forall j \in \{1, 13\}. \tag{26}$$

The similarity is defined as:

$$\mathbf{T} \cdot \mathbf{x} = \mathbf{u} + s\mathbf{R}(\gamma) \cdot \mathbf{x}, \ \forall \mathbf{x} \ \in \mathbb{R}^2, \tag{27}$$

in which $(\mathbf{u}, \gamma, s)$ are offset, rotation angle and scaling factor respectively. These parameters are readily calculated using head center $\mathbf{x}_H$ and "feet" location on the ground floor $\mathbf{x}_F$ in training and testing images.

The results obtained when varying σ_{XY} and σ_θ are given in Fig. 12. The first observation is that for lower level of noise the proposed homographic alignment outperforms the similarity alignment. If a good localization and viewpoint estimation are provided by the trackking algorithm, the pose estimation is more accurate using the projective transformation (3.5 pixels) instead of the similairy transform (4 pixels). The average pose error almost linearly increases with increasing localization noise η_{XY} for both alignment methods, slightly more for the proposed homographic alignment (Fig. 12a). A slight noise in the viewpoint estimation $\sigma_\theta \leq \frac{\pi}{16}$ does not seem to affect any of the 2 alignment methods (Fig. 12b). However, while the error with similarity seems to linearly increase with increasing viewpoint noise η_θ for higher noise levels, the effect is much more pronounced for the projective alignment. By augmenting σ_θ, we slowly increase the possibility of picking the wrong view which has more important con-

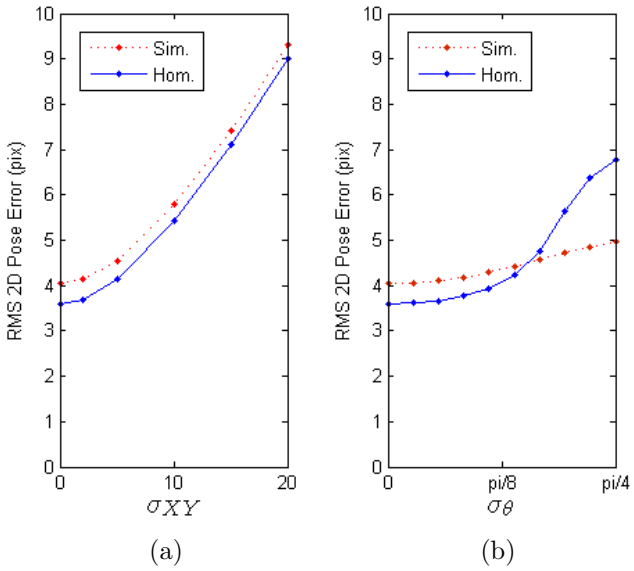


Figure 12: Effect of noise on 2D pose estimation: the average RMS 2D pose error (in pixels) is computed over a set of manually labelled testing poses and a set of training poses aligned using homographic (Hom) and similarity (Sim) alignments. The results are obtained varying the variance of the additive Gaussian white noise which has been added to (a) the ground truth location (X, Y) (in cm) and (b) the viewpoint angle θ (in radians).

sequences when a homography is employed between training and testing view planes instead of a simple Euclidean

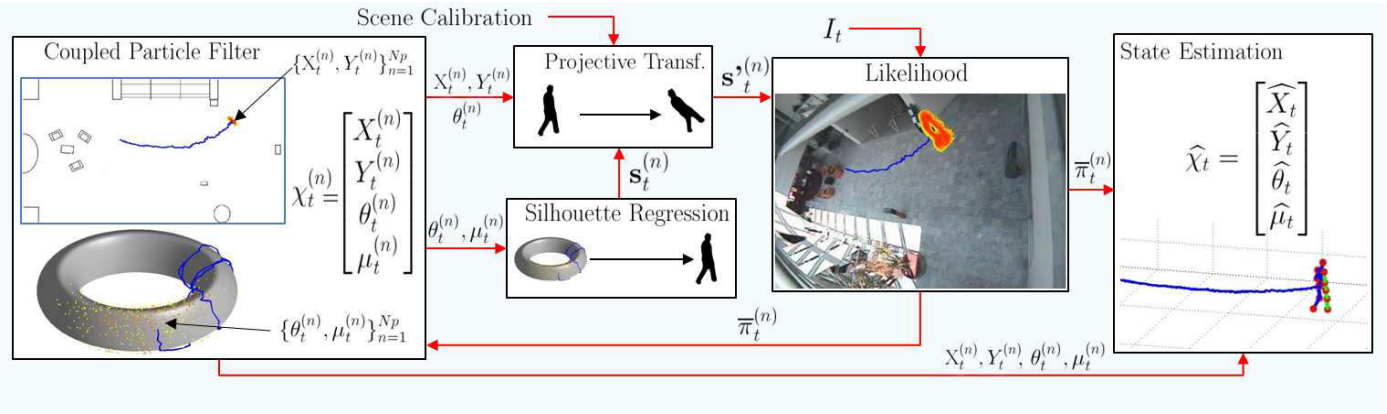


Figure 13: Top-Down System Flowchart: the 3D body poses are tracked using a recursive Bayesian sampling conducted jointly over the scene's ground plane (X, Y) and the pose-viewpoint (θ, μ) torus manifold ([5]). For each sample n, a projective transformation relating the corresponding training plane and the image points is calculated using the dominant 3D directions of the scene, the sampled location on the ground plane $(X_t^{(n)}, Y_t^{(n)})$ and the sampled camera view $\theta_t^{(n)}$. Each regressed silhouette shape $\mathbf{s}_t^{(n)}$ is projected using this homographic transformation obtaining $\mathbf{s}_t'^{(n)}$ which is later matched in the image to estimate its likelihood and consequently the importance weight. A state, i.e., an oriented 3D pose in 3D scene, is then estimated from the sample set.

transformation. The benefit of using a homographic alignment rapidly decreases with the amount of added noise in viewpoint estimation $\sigma_\theta \geq \frac{\pi}{8}$ and the error even gets larger than the one obtained with a similarity transformation for $\sigma_\theta \geq \frac{\pi}{6}$.

4.1.4. Discussion

Acceptable results have been obtained for sequences with a single walking subject but we identified two main drawbacks: 1) estimating the viewpoint from the trajectory does not allow to handle static cases and 2), more importantly, the pose estimation result greatly depends on the accuracy achieved when estimating both location (X, Y) and orientation θ. This first framework performs sufficiently well when an accurate estimation of both ground plane location and orientation (i.e., viewpoint) can be made but, with high levels of noise, the effect on pose estimation is much more pronounced for our proposed projective alignment.

Our numerical experiments show that if a good estimation can be made of both ground plane location (X, Y) and viewpoint θ, the proposed projective transformation outperforms the commonly employed similarity transform. Other types of detector which incorporate an estimation of the orientation θ could be considered in future work to avoid estimating it from the trajectory. For example, multi-class detector such as [13, 14] or head-shoulder detector could be combined with [39] to track people and estimate the orientation θ in perspective video sequences.

4.2. Shape Transformation and Top-down Matching

We now show how our proposed projective transformation can be used to deform a candidate silhouette, in a particular filter paradigm, before its top-down matching in the image. We evaluate this homographic alignment when employed within the tracking framework that we proposed in [48]. In this section, we show that our method works better than the commonly used similarity alignment which was used in [7, 13, 15, 47].

4.2.1. Top-down framework

Fig. 13 shows how the proposed projective transformation fits in a whole human motion analysis framework. A stochastic approach is followed for estimating both location and viewpoint, and the optimum projective transformation for pose recognition is estimated by sampling multiple possible values for θ at multiple locations (X, Y). Applying a different projective transformation to the input image for each sampled triplet $(X^{(n)}, Y^{(n)}, \theta^{(n)})$ and processing each resulting warped image in a *bottom-up* manner as in the previous section would be computationally inefficient. Instead a *top-down* approach is followed where for each triplet, a silhouette $\mathbf{s}_t^{(n)}$ is sampled, transformed using the inverse projection $\mathbf{P}_{I\,\Pi_v\Phi_v} = (\mathbf{P}_{\Phi_v\Pi_v\,I})^{-1}$ and later matched in the original input image. A low dimensional torus manifold for camera viewpoint and pose parameter is used to model 3D walking poses as in [5]. This manifold is mapped to the view-based silhouette manifolds using kernel-based regressors, which are learnt using a Relevance Vector Machine (RVM). Given a point on the surface of the torus, the resulting generative model can regress the corresponding pose and view-based silhouette.

During the online stage, 3D body poses are thus tracked using a recursive Bayesian sampling conducted jointly over the scene's ground plane and this pose-viewpoint torus manifold, in a 4-dimensional state space defined as:

$$\chi_t = [X_t\ Y_t\ \theta_t\ \mu_t\ ], \tag{28}$$

consisting of the ground plane location (X_t, Y_t) and the coordinates on the torus surface $(\mu_t, \theta_t) \in [0,\ 1) \times [-\pi, \pi]$. For each sample n , the homography $\mathbf{P}_{I\,\Pi_v\Phi_v}$ relating the

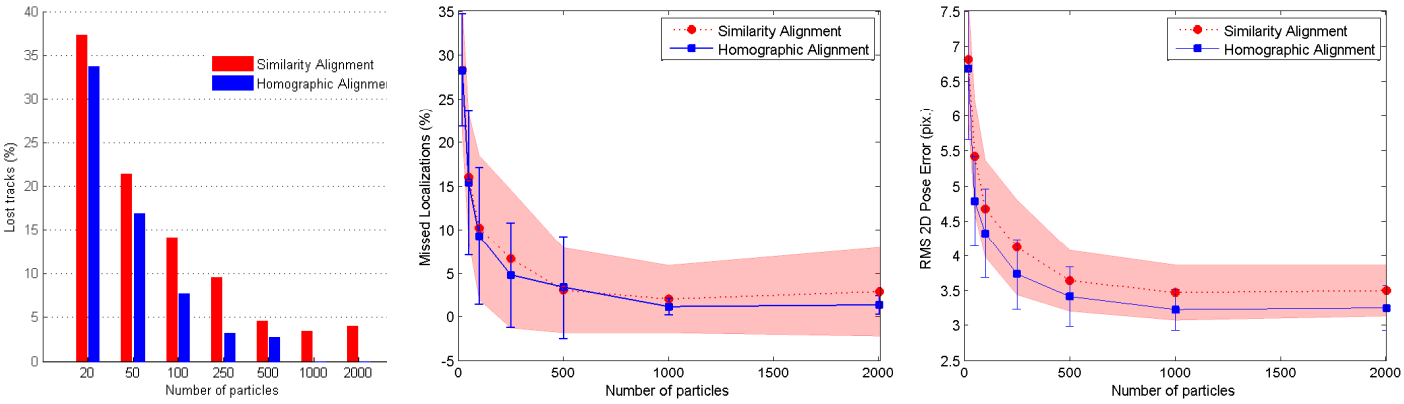


Figure 14: (a) Percentage of lost tracks vs number of particles for similarity and homographic alignment. We present the average performance over 20 runs of the tracking algorithm on the 11 sequences: a track is considered lost when the tracking has failed during 20 frames or more (the distance between the nearest particle and ground truth location is over 1 meter) and it has not recovered by the end of the sequence, i.e., in the last frame the subject is still one meter away from ground truth for the nearest particle. (b) Percentage of missed localizations (mean and std). Again, a localization is valid if the distance between nearest particle and ground truth location is below 1 meter otherwise it is considered to be missed. (c) corresponding average 2D pose error (and std) computed using only valid localizations fom (b). The pose error is computed as the RMSE between nearest particle and ground truth using the 13 2D-joints in pixels.

corresponding training plane to the image points, calculated following Alg. 1, is thus employed to project the associated regressed silhouette shape $\mathbf{s}_t^{(n)}$ obtaining $\mathbf{s}_t'^{(n)}$ which is later matched in the image to estimate its likelihood and consequently the importance weight π_t^n.

The state $\widehat{\chi}_t$ is computed at each time step by using a Monte Carlo approximation of the expectation of the posterior pdf, i.e., a weighted sum over the set of samples: $\widehat{\chi}_t^{MC} = \mathcal{E}[\chi_t] = \sum_{n=1}^{N} \overline{\pi}_t^n \chi_t^n$, where $\overline{\pi}_t^n$ stands for the normalized weight assigned to each particle n. This leads to the estimation of a 2D pose $\widehat{\mathbf{k}}_t$ which is used for numerical evaluation in the next section. To deal with multiple targets, we instantiate several independent 1-subject trackers and model each subject's 3-D occupancy on the ground floor with a Gaussian probability function centered on the subject's estimated location that is employed to downweight the particles from the other targets. The approach proposed in [68] could be employed to deal with more severe occlusions.

4.2.2. Experiments and numerical evaluation

To demonstrate the efficiency of the proposed projective transformation, we process a set of 11 sequences from the Caviar dataset [21] and present a numerical evaluation for $N_{GT} = 2784$ poses $\mathbf{k}_t^{GT}$ which have been manually labelled. More details on this test dataset can be found in [48]. Since randomness is involved in the sampling procedure, to gain statistical significance, we perform the same experiments 20 times and compute numerical result as the average over these 20 runs. We consider that a target has been lost and the localization is not valid if the minimum distance (in the set of particles) to ground truth location exceeds 1 meter, i.e., $\min_{n\in\{1,N\}} \left(||\chi_t^{GT} - \chi_t^{(n)}||_{gd} \right) \geq$ 100 cm, where $||.||_{gd}$ is the Euclidean distance on the ground floor. We believe that a pose estimation does not make sense if the nearest particle is 1 meter away. A track is then considered lost when then the target has been lost during 20 frames or more and has not been recovered in the last frame of the sequence.

Numerical results show that the proposed homographic alignment reduces the average percentage of lost tracks as can be observed in Fig. 14a. The percentage of lost tracks decreases with the number of particles employed in the filter for both methods, but we reach 0% of lost tracks with 1000 particles and over while 5% of the tracks are still lost when considering 2000 particles and a similarity transformation. The perspective correction allows for better shape matching and consequently a more efficient shape-based tracking. If we compute the number of valid localizations, defined as the cases where the distance between the nearest particle and ground truth location is below 1 meter, the tracker loses fewer targets when a homographic alignment is used rather than a similarity alignment (see Fig. 14b). We even reach an average of 99% of valid localizations above 1000 particles.

The pose estimation performances, computed as the RMS distance d_k in pixels (see Eq. 24) between evaluated 2D poses $\widehat{\mathbf{k}}_t$ and ground truth poses $\mathbf{k}_t^{GT}$, are depicted in Fig. 14c. We can observe how the 2D pose error decreases with the number of particles and how the framework, again, performs better when a projective transformation is used and allows for a more accurate pose estimation.

We then carry out a deeper analysis of the different results and compute the different rates in function of the distance between the subject and the camera. In Tab. 2, we present the average percentage of missed localizations,

Table 2: Performances w.r.t. the distance to the camera: percentage of missed localizations (top row), 2D pose error (middle row) and (X, Y) ground plane localization error (bottom row) are given for 20, 50, 100, 250, 500 and 1000 particles. Note that the different values are computed using the poses from 0 meter up to the given distance to the camera D_{max} (5, 10 and 15 meters). Missed localizations from the top row have not been taken into account to compute the performances in middle and bottom rows. Again 2D pose error and (X, Y) ground plane localization error are computed as the RMSE between estimation and ground truth using the 13 2D-joints locations in pixels and the 2D location in cm respectively.

Alignment		**Similarity**						**Homography**					
No. Particles		20	50	100	250	500	1000	20	50	100	250	500	1000
Missed	$D_{max} = 5$	24.30	15.26	8.83	8.59	2.88	1.8	18.02	12.17	7.83	4.92	4.26	0
Loc.	$D_{max} = 10$	30.57	18.79	11.87	8.25	3.73	2.52	29.34	17.23	11.14	5.74	4.17	1.46
(%)	$D_{max} = 15$	28.21	15.97	10.19	6.63	3.05	2.04	28.20	15.31	9.27	4.78	3.39	1.18
Pose	$D_{max} = 5$	9.93	7.32	6.37	5.42	5.18	4.98	9.41	5.58	5.06	4.45	4.30	4.15
Error	$D_{max} = 10$	7.53	5.99	5.02	4.47	3.92	3.72	7.29	5.05	4.56	3.95	3.67	3.40
(pix.)	$D_{max} = 15$	6.80	5.42	4.67	4.12	3.64	3.47	6.67	4.77	4.32	3.74	3.41	3.23
X,Y	$D_{max} = 5$	23.61	16.06	12.23	9.28	7.95	7.26	21.12	9.70	8.21	6.16	5.67	4.86
Error	$D_{max} = 10$	26.7	21.16	16.95	14.17	11.15	10.09	26.22	17.96	15.53	12.76	11.28	10.01
(cm)	$D_{max} = 15$	27.84	22.15	18.55	15.48	12.61	11.61	28.2	20.43	17.9	14.71	12.9	12.09

the average 2D pose error $d_k(\mathbf{k}_t^{GT}, \widehat{\mathbf{k}_t})$ and the average ground plane location error $||\chi_t^{GT} - \widehat{\chi_t}||_{gd}$ varying the size of the particles sets. Results are presented for 3 different maximum distances D_{max} to the camera[7]: 5, 10 and 15 meters. Note that the different values are computed using the poses from 0 meter up to the given distance D_{max}. In the middle row, we can observe that, given a set of particles, the average pose error globally decreases as we augment the maximum distance to the camera D_{max} and add new poses further away. The opposite happens with the ground plane location error (bottom row). This is expected because when people move away from the camera their size in the image gets smaller. Thus, the 2D pose gets smaller when moving away from the camera leading to a consecutive lower 2D pose error while an accurate localization on the ground plane becomes more difficult with the distance. We also want to point out that this numerical evaluation is computed using ground truth data obtained from manual labelling whose accuracy and reliability also decrease with the distance to the camera. The improvement achieved by the proposed homographic alignment is more pronounced when the subjects are close to the camera ($D_{max} = 5$). This makes sense since the viewpoint changes when a subject moves far away from the camera and tends to a tilt angle $\varphi = 0$ which is similar to the training viewpoint employed in this paper.

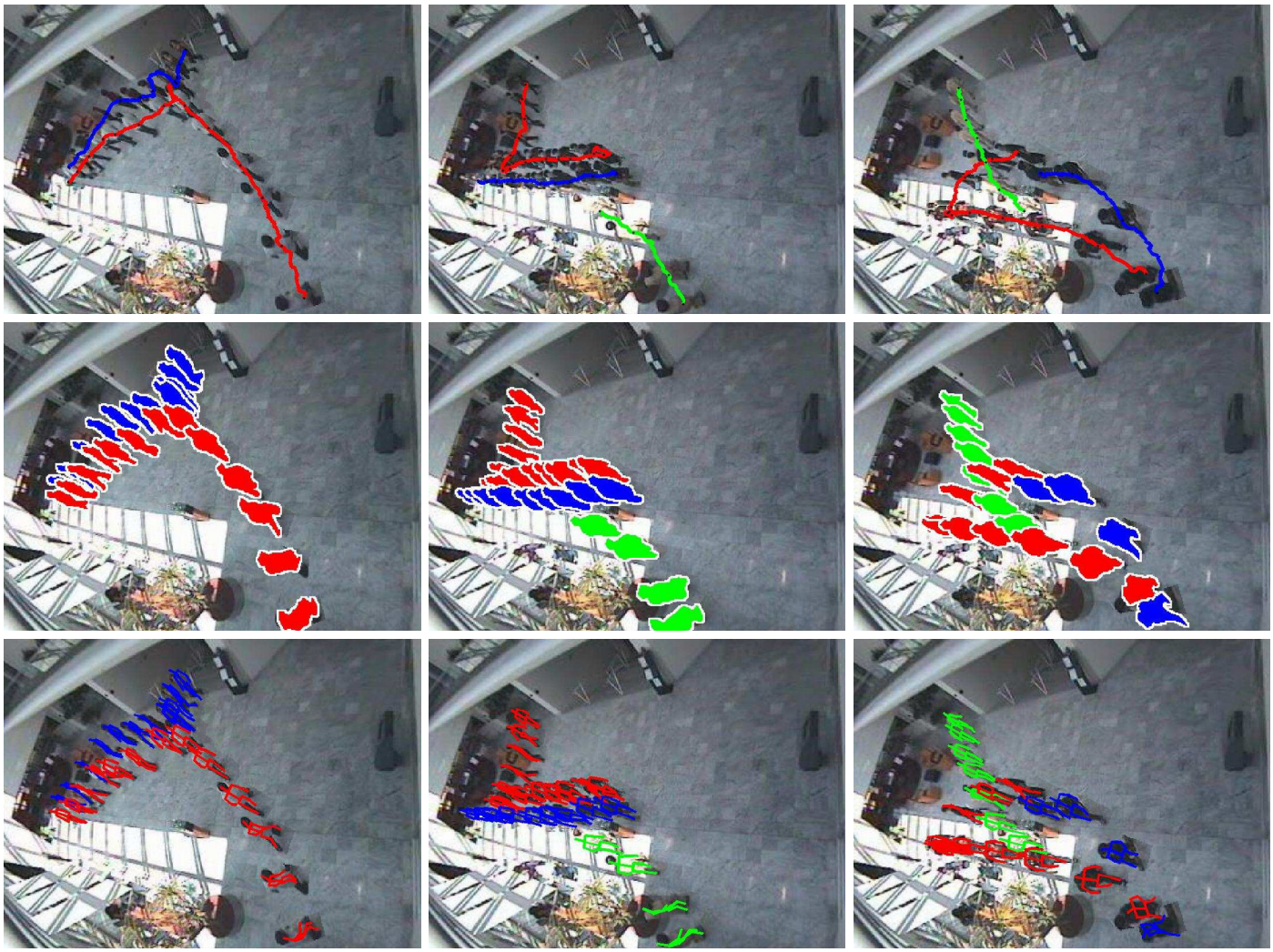
We now present qualitative results using the framework with our proposed homographic alignment and 1000 particles, for 2 complex sequences with multiple interacting subjects in Fig. 15. For each sequence, we can observe the trajectories of the subjects in the image and the tracked silhouettes for a few frames as well as the 2D poses ontop of the image and the estimated 3D poses which have been successfully tracked. We can observe how the system is robust to occlusions and static phases as in both sequences, the subjects walk, meet and stop to talk and then, walk together. Experiments show that the framework, when associated with the proposed homographic alignment, successfully tracks walking pedestrians and estimate their poses in cases where a small number of people move together, have occlusion, and cast shadow.

[7]The distance to the camera D is computed a the Euclidean distance between (X_C, Y_C), the projection on the ground plane of the camera center, and (X^{GT}, Y^{GT}) the ground truth location.

5. Conclusions

In this paper, we have presented a method for view invariant monocular human motion analysis in man-made environments. We have assumed that the camera is calibrated w.r.t. the scene and that observed people move on a known ground plane, which are realistic assumptions in surveillance scenarios. Then, we have proposed to discretize the camera viewpoint into a series of training viewpoints and align input and training images. We have demonstrated that exploiting projective geometry alleviates the problems caused by roof-top and overhead cameras with high tilt angles, and have shown that using 8 training views was enough to produce acceptable results when using the proposed projective alignment in a silhouette-based motion analysis framework.

We have analyzed the results obtained when this homographic transformation is included within two different frameworks: 1) a bottom-up image analysis system where the homography is used to align an input image to a se-



(a) (b) (c)

Figure 15: Qualitative tracking results using our projective method for view-invariant pose tracking and 1000 particles, for the *Meet_WalkTogether2* sequence with 2 interacting subjects in (a) and the *Meet_Split_3rdGuy* sequence with 3 interacting subjects, first (frames 1-330) and second half (frames 330-660) in (b) and (c) respectively. For each sequence, we show from top to bottom: the trajectories in the image, some of the tracked silhouettes, the 2D poses (which have been used for numerical evaluation) and the estimated 3D poses. Results are presented in the attached videos *Meet_WalkTogether2_processed.avi* and *Meet_Split_3rdGuy_processed.avi*.

lected training plane for a view-based processing, and 2) a top-down tracking framework where the inverse homography is employed to transform and project candidate silhouettes in the image.

We have conducted a series of experiments to quantitatively and qualitatively evaluate this projective alignment for a variety of sequences with perspective distortion, some with multiple interacting subjects and occlusions. In our experimental evaluation, we have demonstrated the significant improvements of the proposed projective alignment

over a commonly used similarity alignment and have provided numerical pose tracking results which demonstrate that the incorporation of this perspective correction in the top-down pose tracking framework results in a higher tracking rate and allows for a better estimation of body poses under wide viewpoint variations.

We have also analyzed the limitations of the proposed method in the bottom-up framework by evaluating its sensitivity to noisy measurements. We have observed that the result depends on the accuracy achieved when estimating both location on the ground floor and the orientation of the subject with respect to the camera. This problem does not appear when following a stochastic approach for estimating the optimum projective transformation by sampling multiple possible values for the camera viewpoint at multiple locations. Even if our results show that the top-down framework outperforms the bottom-up one, we believe that bottom-up techniques could benefit from a more sophisticated tracker and might outperform top-down approaches as different types of image features could then be employed.

Even if all the presented experiments are specific to the walking activity (due to the higher availability of training and evaluation datasets), our method is general enough to extend to other activities. The limited number of required training views makes our work easily extendable to more activities and makes more feasible the development of future action recognition software in real surveillance applications.

Acknowledgement

Part of this work was conducted while the first author was a Research Fellow at Oxford Brookes University. This work was also supported by Spanish grants TIN2010-20177, DPI2012-31781, FEDER and by the regional government DGA-FSE. Prof. Torr is in receipt of a Royal Society Wolfson Research Merit Award. Dr Rogez is currently funded by the European Commission under a Marie Curie Fellowship.

References

[1] G. Shakhnarovich, P. Viola, R. Darrell, Fast pose estimation with parameter-sensitive hashing, in: ICCV, 2003.

[2] G. Mori, J. Malik, Recovering 3d human body configurations using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (2006) 1052–1062.

[3] E. Ong, A.S. Micilotta, R. Bowden, A. Hilton, Viewpoint invariant exemplar-based 3d human tracking, Computer Vision and Image Understanding 104 (2006) 178–189.

[4] A. Agarwal, B. Triggs, Recovering 3d human pose from monocular images, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (2006) 44–58.

[5] A.M. Elgammal, C.S. Lee, Tracking people on a torus, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 520–538.

[6] C.S. Lee, A.M. Elgammal, Coupled visual and kinematic manifold models for tracking, International Journal of Computer Vision 87 (2010) 118–139.

[7] T. Jaeggli, E. Koller-Meier, L.J.V. Gool, Learning generative models for multi-activity body pose estimation, International Journal of Computer Vision 83 (2009) 121–134.

[8] C. Sminchisescu, A. Kanaujia, D.N. Metaxas, BM^{3}E : Discriminative density propagation for visual tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2007) 2030–2044.

[9] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: CVPR, Vol. 2, 2005, pp. 886–893.

[10] M. Enzweiler, D.M. Gavrila, Monocular pedestrian detection: Survey and experiments, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 2179–2195.

[11] M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, L.J.V. Gool, Online multiperson tracking-by-detection from a single, uncalibrated camera, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (2011) 1820–1833.

[12] J. Gall, A. Yao, N. Razavi, L.J.V. Gool, V.S. Lempitsky, Hough forests for object detection, tracking, and action recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (2011) 2188–2202.

[13] G. Rogez, J. Rihan, C. Orrite, P.H. Torr, Fast human pose detection using randomized hierarchical cascades of rejectors, International Journal of Computer Vision 99 (2012) 25–52.

[14] R. Okada, S. Soatto, Relevant feature selection for human pose estimation and localization in cluttered images, in: ECCV, 2008, pp. 434–445.

[15] M. Andriluka, S. Roth, B. Schiele, Monocular 3d pose estimation and tracking by detection, in: CVPR, 2010, pp. 623–630.

[16] D. Ramanan, D.A. Forsyth, A. Zisserman, Tracking people by learning their appearance, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2007) 65–81.

[17] B. Wu, R. Nevatia, Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses, International Journal of Computer Vision 82 (2009) 185–204.

[18] L.D. Bourdev, S. Maji, T. Brox, J. Malik, Detecting people using mutually consistent poselet activations, in: ECCV (6), 2010, pp. 168–181.

[19] P.F. Felzenszwalb, R.B. Girshick, D.A. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010) 1627–1645.

[20] Z. Lin, L.S. Davis, Shape-based human detection and segmentation via hierarchical part-template matching, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010) 604–618.

[21] EC funded CAVIAR project IST 2001 37540 (2004).
URL `http://homepages.inf.ed.ac.uk/rbf/CAVIAR/`

[22] R. Gross, J. Shi, The CMU motion of body (MoBo) database (2001).

[23] X. Ji, H. Liu, Advances in view-invariant human motion analysis: A review, IEEE Transactions on Systems, Man, and Cybernetics, Part C 40 (2010) 13–24.

[24] I. Haritaoglu, D. Harwood, L. Davis, W4: Real-time surveillance of people and their activities, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 809–830.

[25] M. Isard, J. MacCormick, BraMBLe: A bayesian multiple-blob tracker, in: ICCV, 2001, pp. 34–41.

[26] N.T. Siebel, S.J. Maybank, Fusion of multiple tracking algorithms for robust people tracking, in: ECCV, 2002, pp. 373–387.

[27] R. Rosales, M. Siddiqui, J. Alon, S. Sclaroff, Estimating 3d body pose using uncalibrated cameras, CVPR 1 (2001) 821–827.

[28] R. Cucchiara, C. Grana, A. Prati, R. Vezzani, Probabilistic posture classification for human-behavior analysis, IEEE Trans. Systems, Man, and Cybernetics - part A 35 (2005) 42–54.

[29] A. Farhadi, M.K. Tabrizi, Learning to recognize activities from the wrong view point, in: ECCV (1), 2008, pp. 154–166.

[30] V. Parameswaran, R. Chellappa, View independent human body pose estimation from a single perspective image, in: CVPR (2), 2004, pp. 16–22.

[31] V. Parameswaran, R. Chellappa, View invariance for human action recognition, International Journal of Computer Vision 66 (2006) 83–101.
[32] A. Datta, Y. Sheikh, T. Kanade, Linearized motion estimation for articulated planes, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (2011) 780–793.
[33] I. Bouchrika, M. Goffredo, J.N. Carter, M.S. Nixon, Covariate analysis for view-point independent gait recognition, in: ICB, 2009, pp. 990–999.
[34] M. Goffredo, R.D. Seely, J.N. Carter, M.S. Nixon, Markerless view independent gait analysis with self-camera calibration, in: FG, 2008, pp. 1–6.
[35] D. Gong, G.G. Medioni, Dynamic manifold warping for view invariant action recognition, in: ICCV, 2011.
[36] K. Grauman, G. Shakhnarovich, T. Darrell, Example-based 3d shape inference from a single silhouettes., in: Proc. ECCV Workshop SMVP, 2004.
[37] A. Kale, A.K.R. Chowdhury, R. Chellappa, Towards a view invariant gait recognition algorithm, in: IEEE Int. Conf on Advanced Video and Signal based Surveillance, 2003, pp. 143–150.
[38] G. Rogez, J. Guerrero, J. Martínez, C. Orrite, Viewpoint independent human motion analysis in man-made environments, in: Proc. of the 17th British Machine Vision Conference (BMVC), Vol. 2, Edinburgh, UK, 2006, pp. 659–668.
[39] Y. Li, B. Wu, R. Nevatia, Human detection by searching in 3d space using camera and scene knowledge, in: ICPR, 2008, pp. 1–5.
[40] J. Gall, B. Rosenhahn, T. Brox, H.P. Seidel, Optimization and filtering for human motion capture, International Journal of Computer Vision 87 (2010) 75–92.
[41] A. Balan, L. Sigal, M. Black, J. Davis, H. Haussecker, Detailed human shape and pose from images, in: CVPR, 2007, pp. 1–8.
[42] R. Rosales, S. Sclaroff, Combining generative and discriminative models in a framework for articulated pose estimation, International Journal of Computer Vision 67 (2006) 251–276.
[43] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, P.H. Torr, Randomized trees for human pose detection, in: Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2008.
[44] X. Lan, D.P. Huttenlocher, A unified spatio-temporal articulated model for tracking., in: CVPR (1), 2004, pp. 722–729.
[45] G. Rogez, C. Orrite, J. Martínez, A spatio-temporal 2d-models framework for human pose recovery in monocular sequences, Pattern Recognition 41 (2008) 2926–2944.
[46] L. Sigal, A.O. Balan, M.J. Black, Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion, International Journal of Computer Vision 87 (2010) 4–27.
[47] K. Toyama, A. Blake, Probabilistic tracking with exemplars in a metric space, International Journal of Computer Vision 48 (2002) 9–19.
[48] G. Rogez, J. Rihan, J.J. Guerrero, C. Orrite, Monocular 3-d gait tracking in surveillance scenes, to appear in IEEE Transactions on Cybernetics.
[49] T. Zhao, R. Nevatia, B. Wu, Segmentation and tracking of multiple humans in crowded environments, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 1198–1211.
[50] D.M. Gavrila, A bayesian, exemplar-based approach to hierarchical shape matching, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2007) 1408–1421.
[51] A. Baumberg, D. Hogg, Learning flexible models from image sequences., in: ECCV, 1994, pp. 299–308.
[52] N.T. Siebel, S.J. Maybank, Fusion of multiple tracking algorithms for robust people tracking, in: ECCV (4), 2002, pp. 373–387.
[53] J. Giebel, D. Gavrila, C. Schnörr, A bayesian framework for multi-cue 3d object tracking, in: ECCV (4), 2004, pp. 241–252.
[54] D. Cremers, Dynamical statistical shape priors for level set-based tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (2006) 1262–1273.
[55] R. Li, T.P. Tian, S. Sclaroff, M.H. Yang, 3d human motion tracking with a coordinated mixture of factor analyzers, International Journal of Computer Vision 87 (2010) 170–190.
[56] D. Weinland, E. Boyer, R. Ronfard, Action recognition from arbitrary views using 3d exemplars, in: ICCV, 2007, pp. 1–7.
[57] M.F. Abdelkader, W. Abd-Almageed, A. Srivastava, R. Chellappa, Silhouette-based gesture and action recognition via modeling trajectories on riemannian shape manifolds, Computer Vision and Image Understanding 115 (2011) 439–455.
[58] T. Riklin-Raviv, N. Kiryati, N.A. Sochen, Prior-based segmentation and shape registration in the presence of perspective distortion, International Journal of Computer Vision 72 (2007) 309–328.
[59] G. Rogez, Advances in monocular exemplar-based human body pose analysis: Modeling, detection and tracking, Ph.D. thesis, Dept. Electron. Comm. Eng., Univ. Zaragoza, Zaragoza, Spain (june 2012).
[60] G. Rogez, J.J. Guerrero, C. Orrite, View-invariant human feature extraction for video-surveillance applications, in: Proc. of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2007, pp. 324–329.
[61] J. Sola, T. Vidal-Calleja, J. Civera, J. Montiel, Impact of landmark parametrization on monocular ekf-slam with points and lines, International Journal of Computer Vision 97 (2012) 339–368.
[62] R.I. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, 2nd Edition, Cambridge University Press, ISBN: 0521540518, 2004.
[63] T. Zhao, R. Nevatia, Tracking multiple humans in complex situations, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2004) 1208–1221.
[64] A. Criminisi, I.D. Reid, A. Zisserman, Single view metrology, International Journal of Computer Vision 40 (2000) 123–148.
[65] E. Lutton, H. Maitre, J. Lopez-K., Contribution to the determination of vanishing points using hough transform, IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (1994) 430–438.
[66] J. Guerrero, C. Sagüés, Robust line matching and estimate of homographies simultaneously, in: Proc. Ib. Conf. on Pattern Recognition and Image Analysis (IbPria), 2003, pp. 297–307.
[67] T. Zhao, R. Nevatia, Stochastic human segmentation from static camera, in: IEEE Workshop on Motion and Video Computing, 2002.
[68] O. Lanz, Approximate bayesian multibody tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (2006) 1436–1449.