View-invariant Human Feature Extraction for Video-surveillance Applications

Grégory Rogez\(^1\), J.J. Guerrero\(^2\) and Carlos Orrite\(^1\)

\(^1\)Computer Vision Lab - \(^2\)Robotics, Perception and Real Time Group
I3A, University of Zaragoza, SPAIN \{-grogez, jguerrer, corrite\}@unizar.es

Abstract

We present a view-invariant human feature extractor (shape+pose) for pedestrian monitoring in man-made environments. Our approach can be divided into 2 steps: firstly, a series of view-based models is built by discretizing the viewpoint with respect to the camera into several training views. During the online stage, the Homography that relates the image points to the closest and most adequate training plane is calculated using the dominant 3D directions. The input image is then warped to this training view and processed using the corresponding view-based model. After model fitting, the inverse transformation is performed on the resulting human features obtaining a segmented silhouette and a 2D pose estimation in the original input image. Experimental results demonstrate our system performs well, independently of the direction of motion, when it is applied to monocular sequences with high perspective effect.

1. Introduction

The potential benefits of an automatic video understanding system in surveillance applications have stimulated the investigation in computer vision, specially in the areas related to human motion analysis. Shape models have shown very encouraging results for Shape-based human detection and tracking [1, 2, 3]. The bi-dimensional nature of image features makes the correspondences with 2D-models much easier to establish than with 3D volumetric models. The main disadvantage of 2D-models is their direct dependence to the point of view: the accuracy of the result depends strongly on the similarity of the input viewpoint with the training one. Roof-top cameras are widely utilized for video surveillance applications. They are usually placed at a significant angle from the floor, introducing perspective effects that deform the human silhouette in a way that traditional 2D-models can not be applied correctly.

The goal of this work is to solve the problems of viewpoint dependency thus segmenting, tracking and estimating pedestrian pose using view-based shape models, independently of the point of view from which the scene is observed. The challenging objective is to construct such 2D-models from few training views and make use of them successfully with all possible sequences taken from a single fixed camera with arbitrary viewing angle. In this paper, a solution is proposed for structured man-made environments: we will considered that the camera model is known and will make the assumption that observed people move on a known ground plane. Moreover, since walking is certainly the most observed activity in surveillance sequences and because of the possible application to gait recognition, we will consider the pose recovery of walking pedestrian.

Related Work. Our work is based on Zhao and Nevatia [4] previous results for detection and tracking in complex situations. The principal difference relies on the type of human model considered and on the way it is fit to the image: while a coarse 3D shape model (an ellipsoid) is used in [4], in our proposal, segmentation, tracking and pose estimation are achieved all together using a more detailed Shape+Structure human model [5, 6, 7] matched independently from the viewpoint. Previous works have attempted to find solutions to this view dependency. In [8], 3D pose is estimated from a set of uncalibrated camera views. In [9], a calibrated approach is used in order to avoid perspective distortion of the extracted features. Grauman [6] presented a method for inferring a 3D shape from a single silhouette by collecting multi-view shape examples and finding the ones most likely to have generated the observed contour. In [10], the viewpoint invariance is achieved by projecting all the images onto the ground plane. In [11], a method is proposed for view invariant gait recognition: a side-view is synthesized by homography, considering a pedestrian walking along a straight line.

1.1 Overview of the Approach.

The position of the camera with respect to the observed object, the view, can be parameterized with two angles, latitude \(\varphi\) and longitude \(\theta\), that define the upper viewing hemisphere (Fig. 1a). Our proposal relies on two separate stages:

The off-line stage consists in discretizing the viewing hemisphere into a finite number of training viewpoints (Fig. 1b) and construct a framework of view-based 2D-models combining Shape and 2D-Pose parameters (Fig. 1c).

The on-line stage is considered when processing a video-surveillance sequence with arbitrary viewpoint. Ba-
sically, it consists in selecting the adequate training view, then establish viewpoint-correspondences by projecting the input image onto this training plane and finally employ the selected view-based models for feature extraction in the warped image.

A quite classical global process of Detection-Segmentation-Tracking [4] is considered during on-line stage. The novelty appear in the Human Segmentation block (Fig. 2): people are tracked on the ground plane and an estimation is made of their orientation with respect to the camera thus allowing the selection of the adequate training view; the projection of the input image onto the plane of corresponding training images then enables the compensation of both discretization along $\theta$ and variations along $\phi$. This transformation is achieved using the dominant 3D directions of the scene in both training and input images. Once human features have been evaluated in the warped image, they are back-projected on the original image.

2. View-based Shape-Skeleton Models

As in [3], we extracted precise training human Shapes from CMU MoBo dataset considering 8 different viewpoints, being $\theta$ uniformly distributed between 0 and $2\pi$ and $\phi \approx 0$. Simultaneously, we labelled 13 fundamental 2D-points corresponding to a Stick model: $k_i = \{u_{k,i}, v_{k,i}\}_{j=1}^{13}$. The training Shapes are normalized to 100 points: $s_i = \{u_{s,i}, v_{s,i}\}_{j=1}^{100}$. Figure 1c shows the 8 view-based Shapes & Skeletons for a particular training snapshot: the frontal ($F$), the 2 diagonal ($D_1$ & $D_2$), the 2 lateral ($L_1$ & $L_2$), the 2 rear diagonal ($RD_1$ & $RD_2$) and the back view ($B$).

Following [5], a pedestrian model is constructed encapsulating 2D-Shape landmarks and 2D-Skeleton joints. This Shape-Skeleton database is clustered following spatial and temporal criteria: the spatial clustering is directly provided by the 8 training views and the temporal one is obtained by dividing the gait cycles into several gait states. As in [7], a framework of spatio-temporal 2D-models is learnt by fitting a mixture of PCA models to the clustered feature space and a Probabilistic Transition Matrix (PTM) is evaluated frame to frame. This PTM provides the transition probability imposing spatial and temporal constraints and limits the feature space to the most probable models of the framework. More complete explanations about the framework construction and the constraint-based search can be found in [7].

3 Solving Viewpoint Dependency

As demonstrated in [11], 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 far enough from the camera can be approximated by a plane. Note that in the following sub-sections upper case letters ($X$) will be used to indicate quantities in space whereas image quantities will be indicated with lower case letters ($x$).

3.1 Projection to Vertical Plane

Following the classical notation of 3D projective geometry, a 3D point $[X, Y, Z]^T$ is related to its 2D image projection $[u, v]^T$ via a $3 \times 4$ projection matrix $M$:

$$[u, v]^T = M \cdot [X, Y, Z]^T,$$

where “$\approx$” means equality up to scale. The projective transformation matrix $M$ can be determined with a series of intrinsic and extrinsic parameters or, as shown in [12], it can
be defined in function of the vanishing points of the dominant 3D directions. Suppose we want to transform the image to a frontal view defined as the vertical plane $\pi$ parallel to the horizontal direction $D$, (1) becomes:

$$[u, v, 1]^T = H \cdot [D, Z, 1]^T,$$

with $D$ coordinate on the $D$-axis and $H$ homography matrix defined as:

$$H = [v_D \, \alpha v_Z \, o],$$

where $v_Z$ is the vertical vanishing point, $o$ is the origin of the world coordinate system and $\alpha$ is a scale factor. $v_D$ is the intersection of $d$, projection of $D$ in the image, and $v_L$, the horizontal vanishing line.

Given a direction $d$ on the ground plane, it is back-projected onto the input image leading to the direct localization of $v_D$ and the computation of $H$. Scaling parameter $\alpha$ is evaluated considering 4 points that define a Bounding Box and whose positions in Real World are known. It is computed so that the height-width ratio still presents the same value in the warped image.

In [13], we proposed to project both model and input image to a common frontal view, parallel or orthogonal to the direction of motion, and do the shape registration in this vertical plane. This strategy gave some preliminary encouraging results, but it can be improved by considering more possible viewpoints and a direct projection of the input image onto the model plane thus loosing less information.

### 3.2 Projection to the Best Training Plane

The transformation between input and training images through a vertical plane centered in the human body, can be obtained as the product of 2 homographies. The first one, $H_r^{-1}$, projects the 2D input image points to the vertical plane and the other one, $H_M$, relates this frontal plane to the model plane of viewpoint $r$, as depicted in Fig. 3 (up. left).

We thus obtain the following equation that relates input image points $[u, v, 1]^T$ with training image ones $[U, V, 1]^T$:

$$[U, V, 1]^T = H_M \cdot H_r^{-1} \cdot [u, v, 1]^T.$$

The training views considered in this work (Fig. 1c) are not exactly frontal, that’s why $H_M$ is taken into account. The 8 Homographies $\{H_M\}_{r=1}^{8}$ are computed during the off-line stage and stored for online use.

The angle $\theta$ of the subject’s orientation (direction of walk) w.r.t the camera, is defined on the ground as:

$$\theta = \hat{CV},$$

where $C$ and $V$ are the projections on the ground plane of the camera viewing direction (straight line connecting ob-
ject and camera, pointing out the camera) and the orientation vector as can be observed in Fig. 3.

Table 1: Training View Selection

<table>
<thead>
<tr>
<th>r</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta^r$</td>
<td>$\pi$</td>
<td>$\frac{\pi}{2}$</td>
<td>$\frac{3\pi}{4}$</td>
<td>$\pi$</td>
<td>$\frac{3\pi}{4}$</td>
<td>$\frac{2\pi}{3}$</td>
<td>$\frac{\pi}{2}$</td>
<td>$0$</td>
</tr>
</tbody>
</table>

The problem now consists in selecting the adequate training view $r$ and consequently the adequate $H_r^{g'}$. This view is defined as the one that minimizes the deformation and the information loss. Being $\{\theta^r = C^r V\}_{r=1}^8$ the 8 training values for $\theta$ (see Tab. 1) and given an estimation of $\theta$ of the subject, $r$ is selected so that:

$$r = \arg \min_{r \in [1, 8]} (\theta - \theta^r).$$

(6)

For each training view $r$, the ground axis defining the nearest vertical plane, $D^r$, and $C^r$ are orthogonal. $D^r$ is then related to $V$ so that:

$$\hat{V}D^r = C^r D^r - C^r V = \frac{\pi}{2} - \theta^r.$$

(7)

Once the view has been selecting, $D = D^r$ is computed in the ground plane by rotating $V$:

$$D = R(\frac{\pi}{2} - \theta^r) \cdot V,$$

(8)

where $R$ is a rotation matrix. $D$ is then back-projected to the input image leading to the computation of $d$, the localization of $v_2$ and the computation of $H_r^{-1}$ that relates input image with the selected vertical plane.

Head and feet localization leads to the extraction of a Region of Interest in the image: $I_R$, $H^{-1}_r$, combined with the right model homography $H_{rg}$, is then applied on $I_R$, obtaining $I_{b,x,y}$ that will be used to fit the selected view-based model. The reliability of the warping, and consequently the reliability of the feature extraction, depends strongly on the precision with which both ground plane position $(X, Y)$ and orientation $\theta$ are estimated.

## 4 Human Feature Extraction

### 4.1 Tracking People by their head

A simple but effective way of locating people in an image relies on detecting their head. First of all, the head is the easiest human feature to detect because of the low variability of its shape and its top position in the body. Moreover, in a sequence taken by a rooftop camera, the head is the most visible feature since it is less likely to be occluded. Finally, in a calibrated environment, a good estimation of the ground plane position $(X, Y)$ can be obtained by projecting vertically the head position.

Many papers have proposed to compute the vertical histogram of the foreground blob and scan it for searching peaks as possible head candidates [2, 4]. The problem of this method is that it cannot detect the heads in the interior of the blob. In [14], the authors extend this head candidates search by the use of a head-shoulder model. Following a similar approach, we train such a model by considering only the upper landmarks of our training shapes and learn a mixture of linear head shape models.

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 $v_2$. We also compute the feet candidates (local minima) and the corresponding probable head location. 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. In this way, non-human blobs resulting from shadows and reflections are dismissed.

Considering that the vertical vanishing point $v_2$, the image position of the head $x_h$ and the feet position in the image $x_f$ are collinear, we can compute on this vertical line the following projective-invariant cross-ratio [12]:

$$Cr = \frac{d(x_h, v_2) \cdot d(x_f, x_h)}{d(x_h, x_f) \cdot d(x_f, v_2)},$$

(9)

where $x_f$ is the intersection of the horizon line $v_L$ and the vertical line passing through $x_h$ and $x_f$, and $d(x_1, x_2)$ denotes the distance between $x_1$ and $x_2$. If $Cr$ is known, $x_f$ can be determined from $x_h$ or $x_h$ from $x_f$ using (9).

During the scene calibration the homography matrix $H_g$ is computed. It characterizes the mapping between image and ground plane. The system is initialized in the first frames, estimating $x_f$ (and the cross-ratio $Cr$) by a rough fitting of our model considering various possible heights as in [14]. A tracking is then applied, the state of the each pedestrian (ground position $(X, Y)$) being estimated at each time step using a Kalman filter as in [4]. The tracking process is detailed in Alg.1. Figure 4 shows an example of head-based tracking. $X$, $Y$ and $\theta$ are then estimated frame to frame, thus allowing the selection of training view $r$ in the Human Feature Extraction block (Fig. 3). In case of occlusions between various pedestrians, their ground plane positions permit to process them in the front-to-back order.

### 4.2 Segmenting and Estimating People Pose

We now explain how the model is applied for a joint segmentation and pose estimation of a pedestrian. Two exam-
**Algorithm 1**: Human tracking based on head detector

1. \( (\hat{X}, \hat{Y}) \) is predicted by the filter.
2. \( \hat{x}_f = H_g^{-1} \cdot [\hat{X}, \hat{Y}, 1]^T \).
3. \( \hat{x}_h \) calculated from \( \hat{x}_f \) using (9).
4. Head shape model fitted providing \( x_h \).
5. \( x_f \) calculated from \( x_h \) using (9).
6. Filter parameters updated using \( [X, Y, 1]^T = H_g \cdot x_f \) and \( \theta \) evaluation from the ground plane trajectory.
7. \( \theta, X \) and \( Y \) sent to the Human Segmentation block.

---

**Figure 4**: View invariant tracking based on head detector: Ground Plane trajectory of Walk3 sequence [15] extracted using the head-based tracker. An example with multiple pedestrians is proposed in lower left.

---

**Figure 5**: Two Examples of Features Extraction. The original image (a) is warped to the corresponding plane (b). The blob information (c) is combined with detected edges (d) to select only the valuable image features (e). Shape and image features are then registered along the shape normals (f). This leads to the estimation of both Shape and Skeleton in the projected image (g), that are back-projected onto the original image(h).

---

**5 Preliminary Results**

For the evaluation of the models, a series of sequences of interest has been selected from the Caviar project database [15]: in these sequences people are walking in various directions and the changing perspective effect can be noted. The different walk sequences available are chosen because of the nature of our “pedestrian” model: the model would not be able to analyze other kinds of motion.

In this paper, the transformation matrices are calculated online using the vanishing points computed off-line after manual selection of the parallel lines present in the image.

In Fig.6, we present the features that have been automatically extracted from the sequence processed in Fig.4: we can observe how the direction of motion slowly changes along the sequence and how the images are projected to the adequate model plane. The resulting features are evidently not perfect but, given the complexity of the task (low resolution, high perspective effect), we consider that they are quite acceptable. Some improvements will be proposed for future work.

**6 Conclusions**

The proposed approach for solving viewpoint dependency can be applied to every possible 2D-models. The viewing using Dynamic Programming. The solution results in a smooth contour that is iteratively deformed. When presented a new Shape, we treat the unknown 2D structure as missing variables and estimate both contour and structure parameters with the selected view-based Shape-Skeleton model, as explained in [7] (Fig.5g).
The method has obviously its limitations: the transformation is not always possible and the top-view case is the extreme one. Even so, we have demonstrated that it is sufficient to model only 8 viewing angles for obtaining very interesting preliminary results. In future work, a stochastic approach for viewpoint estimation will improve the search of the best warping and will solve some difficulties such as stationary cases or partial occlusions. Numerical results will validate completely the proposal. Finally, an automatic method for vanishing points detection and homographies computation could be considered to make the system completely automatic.

Acknowledgment

This work is supported by spanish grant TIN2006-11044 (MEyC) and FEDER. Greg Rogez is funded by the Spanish Ministry of Education under FPU grant AP2003-2257.

References