Unsupervised Learning of Human Perspective Context Using ME-DT for Efficient Human Detection in Surveillance

Liyuan Li
Institute for Infocomm Research, Singapore
lyli@i2r.a-star.edu.sg

Maylor K.H. Leung
Nanyang Technological University, Singapore
asmkleung@ntu.edu.sg

Abstract

A novel and automated technique for learning human perspective context (HPC) from a scene is proposed in this paper. It is found that two models are required to describe HPC for camera tilt angle ranging from 0° to 50°. From a scene, the tilt angle can be inferred from the observed human shapes and head/foot positions. Afterward, a novel ME-DT (Model Estimation - Data Tuning) algorithm is proposed to learn human perspective context from live data of various degrees of uncertainties. The uncertainties may come from the variations of human individual heights and poses, and segmentation/recognition errors. ME-DT not only estimates the model parameters from the training data but also tunes the data to achieve a better head-foot correlation. The human perspective context provides a feasible constraint on the scales, positions, and orientations of humans in the scene. Applying this constraint to the HOG human detection, great reduction of the detection windows and improved performances have been obtained compared to conventional methods.

1. Introduction

In video surveillance, a scene is usually monitored by a stationary camera from a high position with an oblique angle to the ground plane. The objects of interest, e.g. humans and vehicles, move on the ground surface. The perspective projection transformation determines the appearances of the target objects in the scene. The knowledge of the perspective about the target objects on the ground plane can not only provide the hints about positions, scales, orientations, and shapes of the target objects in the image but also specify the mapping between the image measurements and physical measurements for the target objects. Hence, the perspective context of a scene is much helpful for object detection, tracking, and event understanding [16, 6].

In principle, the perspective projection of a scene is precisely described by the camera parameters and 3D world geometry of the scene. The approaches to establish the camera parameters are known as camera calibration. The standard methods assume that a special calibration object, long parallel lines, or measurements of enough 3D points in the scene are available [5, 15, 1, 2]. Unfortunately, such manual techniques are not adequate for automated video surveillance systems with widespread deployed cameras.

Recently, a few self-calibration methods from pedestrians have been proposed. These methods were originated from [1] which showed that the vanishing points of parallel lines in the scene can be used to recover camera parameters. Lv *et al*. [10] proposed to use a tracking approach to select samples from a walking person. Linking the points of head and foot for a pair of human samples, two parallel lines in the scene are obtained. From these samples, the vertical vanishing point and vanishing horizon line can be estimated, which are in turn used to calculate the camera parameters. Since both horizontal and vertical vanishing points could be obtained from a pair of samples of a tracked person, in [7], two harmonic homologies were introduced to derive linear equations for the focal length of the camera. Outliers are removed by using Rayleigh quotient. Considering that tracking persons over various occlusions in natural scenes is still a very challenging problem, the methods depending on tracking normally need training sequences which contain a single person walking in the scene. They can be considered as semi-automatic methods. Krahnstoever and Mendonca [8] proposed a fully automatic self-calibration method from detected humans. Foot-to-head homology is first estimated and then decomposed to extract the vanishing point and horizon for calibration. The camera parameters are obtained from a MAP (maximum a posteriori) estimation under the Bayesian framework. Metropolis sampling for random optimization is used to solve the MAP problem. Prior knowledge about camera parameters and sample distributions are needed. Stochastic optimization is robust to noisy data, but it requires not only intensive computations but also good initial estimates.

While camera parameters are usually used in the formal description of perspective projection, it is difficult to

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

Method	Learning	Assump.	Angle	Estimation	Measurement	Tracking	Robustness
Lv [10]	semi-auto	yes	5°-30°	non-linear	real	yes	weak
Junejo [7]	semi-auto	yes	10°-25°	linear	real	yes	strong
Krahustoever [8]	auto	yes	9°-20°	non-linear	real	no	strong
Ours	auto	no	0°-50°	linear	scale	no	very strong

Table 1. The comparison of our method with existing self-calibration methods from captured standing and walking humans.

estimate them from automatically selected observations in natural scenes. In addition, the assumptions on camera parameters for self-calibration might introduce errors for real cameras and scenes. In this paper, we propose to learn and use human perspective context (HPC) of a scene (up to a scale factor of the average human height) to describe human appearances in the scene. It is found that two linear models, i.e. foot⇌head homologies and scale⇌position mappings, are required to describe HPC for camera tilt angles from 0° to 50° to the ground plane. We propose to classify the camera tilt angles into three categories, i.e. large, small and zero angles. When the camera tilt angle varies from large to small, the vertical range of human head positions in the image shrinks faster then that of the foot. An approach is proposed to estimate tilt angle from the statistics of observed human shapes and head/foot positions in images. Human samples are then selected based on the human shape model for the tilt angle. The automatically selected samples can be noisy and uncertain due to the variations of human individual heights and poses, and errors of foreground segmentation and shape recognition. A ME-DT (model estimation - data tuning) algorithm is proposed to learn HPC from such training data. The ME-DT algorithm can not only estimate the model parameters from the training data but also tune the data to achieve a good head-foot correlation. Mathematic evaluation on synthetic data shows that ME-DT algorithm is much more robust than the strategy of simply removing outliers in [7, 8]. The algorithm is implemented for the cases of three tilt angles. The ME-DT algorithm can generate good estimation of HPC from the data set while the conventional linear estimation fails. The comparison of our method to the existing methods for self-calibration from humans is illustrated in Table 1.

The learned HPC is then applied for human detection in the scene. A scene-adaptive grid is generated according to the HPC. At each grid position, the HOG detection window of corresponding scale and orientation determined by the HPC is applied. It is observed that less than 1% of detection windows is required as compared with conventional human detection methods.

The main contribution of this paper is the investigation of a novel and practical approach to learn human perspective context automatically. It includes: (1) a novel approach to estimate camera tilt angle from observed human shapes and distributions of head and foot positions; (2) a new systematic modeling of HPC in surveillance; (3) a ME-DT (model estimation - data tuning) algorithm to estimate HPC from uncertain and noisy data; (4) an efficient method for human detection using HPC. They are described in Sections 2 to 5 in the remaining of the paper. Experiments and conclusions are given in Sections 6 and 7.

2. Camera Tilt Angle Estimation

The scenes observed from three typical tilt angles are illustrated in Figure 1. When observed from a large tilt angle at a high position, the vertical ranges of both head and foot positions of a person in the image are large if the person moves from close to far away positions in the scene. Meanwhile, the body orientations near the left and right margins of the image differ obviously due to the effect of perspective. These visual clues can be observed easily even from noisy samples. On the other hand, when the camera is lowered and have a small tilt angle to the scene, the vertical range of head positions in the image becomes much less than that of the corresponding foot positions while the variations of body orientations caused by the effect of perspective becomes indistinguishable.

According to these observations, in this paper, the camera tilt angles are classified into three categories, *i.e.large*, *small*, and *zero* angles, as depicted in Figure 1. The angular ranges are set empirically as [50°,25°], [25°,8°], and [8°,0°], respectively. Two related sets of visual clues are exploited to estimate the camera tilt angle. One is from the shapes of the observed human individuals while the other is from the vertical distributions of head and foot positions of humans in images.

2.1. Visual Clues from Shape

When observed from different camera tilt angles, the silhouettes of a standing and walking person are different, especially in aspect ratios. On the other hand, when observed from a fixed camera tilt angle, the human silhouettes may vary due to the view angles to the body (*e.g.* front view or side view). To estimate the camera tilt angle from the silhouettes of observed humans, the shape feature needs to distinguish human shape variations caused by the change of the camera tilt angle from those caused by the variations of human poses and view angles to the bodies. Existing human shape descriptors are designed to distinguish humans from

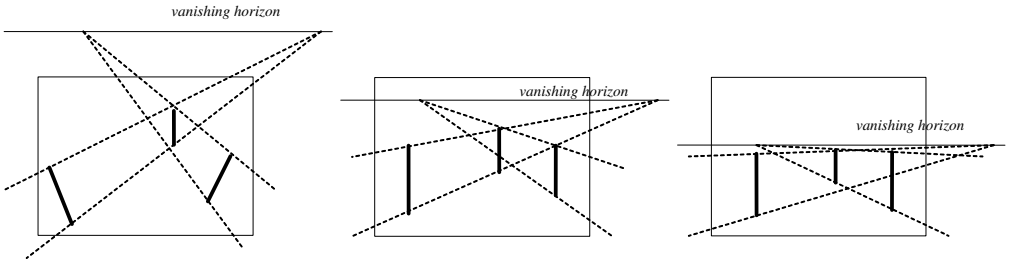

Figure 1. The images from left to right illustrate the cases for large, small, and nearly zero tilt angles of the camera to the scene.

non-human objects. Hence, they cannot be applied.

In this work, three shape features are employed. Firstly, the moment-based descriptor can capture the shape features of standing and walking persons [12]. The head-shoulder width ratio can then be used to distinguish side view from front view. In a front view, the head-shoulder width ratio is small ($\approx 1/3$), but it becomes large (≈ 1) for a side view image. Lastly, the aspect ratio of an observed human can reflect the camera tilt angles.

The 2nd-order moments which is invariant to translation and scaling [9] are employed to describe the shapes of isolated standing and walking persons. Given a silhouette region $R(\mathbf{x})$ with $\mathbf{x} = [x, y]^\tau$ being a pixel in the image, a set of central moments are computed as

$$\mu_{jk} = \sum_{(x,y)} (x - \bar{x})^j (y - \bar{y})^k R(x, y) \qquad (1)$$

where $[\bar{x}, \bar{y}]^\tau$ is the gravity center of the shape. Using (1), the normalized 2nd-order central moments are defined as $\phi_1 = \sqrt{|\mu_{11}|}/|R|$, $\phi_2 = \sqrt{\mu_{20}}/|R|$, $\phi_3 = \sqrt{\mu_{02}}/|R|$, where $|R|$ is the size of the region.

The aspect ratio and head-shoulder width ratio are obtained from the horizontal and vertical projections of the human silhouette. From the projection of R on the horizontal axis, we can obtain the left and right ends x_l and x_r, and from the projection on the vertical axis, we can get the top and bottom ends y_t and y_b. To be robust to noise and shape details, the tails on both sides of a histogram are cut at 10% of the peak height. The aspect ratio of the 2D shape is computed as $r_a = (x_r - x_l)/(y_u - y_l)$. For a standing person, the neck position is usually at 81% location of the height from the foot [4]. The average widths of head and shoulder (w_h and w_s) can be computed from the silhouette parts over and below the neck position. Then the head-shoulder width ratio is obtained as $r_{hs} = w_h/w_s$. The feature vector from an individual shape is defined as $\mathbf{v} = [r_a, r_{hs}, \phi_1, \phi_2, \phi_3]^\tau$.

In this investigation, it is found that the shape features are similar for humans observed from small camera tilt angles of categories 2 and 3. The measurements from category 1, however, are significantly different. Hence, human shapes are firstly categorized into two classes corresponding to *large* and *small* (including *zero*) camera tilt angles divided at about 25°. A multivariate Gaussian is used for each class. The multivariate Gaussian models (*i.e.* $\mathcal{N}_i(\bar{\mathbf{v}}_i, \Sigma_i),\ i = 1, 2$) can be obtained from offline training, where $\bar{\mathbf{v}}_i$ is the mean vector and Σ_i is the covariance matrix. They are then used for online tasks. For a human silhouette R represented by feature vector $\mathbf{v}$, the likelihood of R being observed from a camera tilt angle of the ith class is

$$P_i(\mathbf{v}) \propto \frac{1}{|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{v} - \bar{\mathbf{v}}_i)^\tau \Sigma_i^{-1} (\mathbf{v} - \bar{\mathbf{v}}_i)\right] \qquad (2)$$

The shape features may be affected by camera aspect ratio. Evaluation on sequences from different cameras shows that, with a high threshold, correct classification rates are over 73% and mis-classification rates are below 19% for both models. This result is good enough for the classification based on voting from a large number of samples.

2.2. Visual Clues from Head/Foot Positions

As illustrated in Figure 1, in a scene observed from a large camera tilt angle, the vertical positions of both head and foot vary greatly in the images when a person moves from close to far away positions. However, when the camera tilt angle decreases with the drop of camera position, the vertical extent of head positions shrinks quickly. The comparison between vertical extents of head and foot positions from a large number of observed humans provides a global-level clue about the camera tilt angle.

When a large number of isolated humans around the scene have been observed, the 1D histograms of vertical head and foot positions can be generated. Truncating the 10% tails of a histogram on both sides, the vertical ranges of the head and foot positions can be obtained as d_h and d_f, respectively. The ratio of them is $r_{hf} = d_h/d_f$. In this paper, the linear fuzzy membership is employed to describe the likelihood of camera tilt angle from the ratio value r_{hf}. The membership functions for the 3 tilt angle categories, *i.e.* $Q_j, j = 1, 2, 3$ for *large*, *small*, and *zero* camera tilt angles, are depicted in Figure 2, where the parameters are determined empirically.

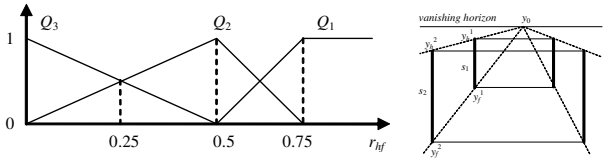

Figure 2. Left: Fuzzy membership functions for the likelihoods of camera tilt angle categories from the ratio r_{hf}. Right: Perspective scene observed from a tilt angle close to 0°.

2.3. Tilt Angle Estimation

Fusing the visual clues from human shapes and head/foot positions, the camera tilt angle can be inferred online. When a surveillance camera is installed, the moving foreground objects can be detected continually from the stream of incoming images using background subtraction. Employing the human shape model, we can select samples of isolated standing and walking persons from these foreground objects. Let $R_t^j(\mathbf{x})$ be the jth foreground object detected at time t and $\mathbf{v}_j$ be its shape feature vector. If it is recognized as a standing or walking person for one of the two categories of camera tilt angles, *i.e.* $\exists i, P_i(\mathbf{v}_j) \geq T_p$, it is selected as one sample of human objects. A large threshold T_p can be used to get good samples since abundant samples are available from a scene.

When enough number (*e.g.* M) of human samples have been collected, the likelihood for the jth category of camera tilt angles from visual clue of human shapes is calculated as $L_j = \frac{1}{M}\sum_m P_j(\mathbf{v}_m)$, where $j = 1, 2, 3$ and $P_3() = P_2()$. The likelihood for the jth category of camera tilt angles from visual clue of head/foot positions of the M samples can be obtained as Q_j. Then, the category of the camera tilt angle is estimated as

$$a = \arg \max_{j\in[1,2,3]} \{L_j + Q_j\} \tag{3}$$

With the estimated camera tilt angle, the set of M samples are further refined, *i.e.* only the samples with $P_a(\mathbf{v}_j) \geq T_p$ are remained. The rest N ($N \leq M$) samples are used to learn human perspective context of the scene.

3. Human Perspective Context

Human perspective context (HPC) is the knowledge of perspective projection of humans in the scene. It determines the scales and orientations of humans at different positions in the image. In principle, the perspective projection determines the transformation from a world coordinate system to the image coordinate frame. Using homogeneous coordinates, let $\tilde{\mathbf{X}} = [X, Y, Z, 1]^\tau$ be a point in a real world coordinate system. The projection transformation is described as $\tilde{\mathbf{x}} = H\tilde{\mathbf{X}}$, where $\tilde{\mathbf{x}} = [\tilde{x}, \tilde{y}, z]^\tau$. The image position is obtained as $x = \tilde{x}/z$ and $y = \tilde{y}/z$.

Let the XY plane be the ground surface. The foot position of a person can be denoted as $\mathbf{X}_f = [X, Y, 0]^\tau$. Since $Z = 0$, one can remove the 3rd column in H. The perspective projection becomes $\tilde{\mathbf{x}}_f = H_f\tilde{\mathbf{X}}_f$ with $\tilde{\mathbf{X}}_f = [X, Y, 1]^\tau$. Similar, if the XY plane is translated upwards to the head plane, there is $\tilde{\mathbf{x}}_h = H_h\tilde{\mathbf{X}}_h$. Both H_f and H_h are 3×3 matrixes. For a standing person, the foot location on the ground plane corresponds to exactly one location in the head plane, hence we have $\tilde{\mathbf{X}}_f = \tilde{\mathbf{X}}_h$ and

$$\tilde{\mathbf{x}}_h = H_{hf}\tilde{\mathbf{x}}_f \text{ and } \tilde{\mathbf{x}}_f = H_{fh}\tilde{\mathbf{x}}_h \tag{4}$$

where $H_{hf} = H_h H_f^{-1}$ and $H_{fh} = H_f H_h^{-1}$. They are 3×3 matrixes and $H_{hf}H_{fh} = H_{fh}H_{hf} = I$. The linear mapping H_{hf} or H_{fh} is called the homology between two parallel planes in real world [11]. The homology can also be derived from the plane vanishing horizon, the vertical vanishing point, and the physical distance between the two planes in the scene [2].

In principle, the foot$\rightleftharpoons$head homologies exist except when the camera is placed at the height of head plane. In this case, the head plane is projected into the vanishing horizon so that H_h becomes singular. Since both H_{hf} and H_{fh} involve H_h, the foot$\rightleftharpoons$head homologies become invalid. In practice, when the tilt angle is close to 0°, it is difficult to obtain a valid homology from noisy data since it is nearly singular. Unfortunately, this case happens frequently when a camera is installed almost horizontally in order to have a large depth coverage. In this case, a simple linear model of scale-position mapping can be used instead. When the camera tilt angle is close to 0° (*e.g.* below 10°), the upright humans are nearly parallel to the image plane, as illustrated in the right of Figure 2. Let y_0 be the vertical location of the vanishing horizon, and s_1 and s_2 be the heights of a person standing at positions y_f^1 and y_f^2 in the image. The corresponding head positions are y_h^1 and y_h^2. According to the relations for similar triangles $\triangle y_0 y_f^1 y_h^1$ and $\triangle y_0 y_f^2 y_h^2$, one can obtain $y_f = k_s s + y_0$, where s is the scale (height) and y_f is the corresponding vertical foot position in the image. The scale$\rightleftharpoons$position mappings can be expressed as

$$s = \mathbf{m}_{sf}\tilde{\mathbf{y}}_f \text{ and } y_f = \mathbf{m}_{fs}\tilde{\mathbf{s}} \tag{5}$$

where $\tilde{\mathbf{s}} = [s, 1]^\tau$ and $\tilde{\mathbf{y}}_f = [y_f, 1]^\tau$.

The two models, *i.e.* foot$\rightleftharpoons$head homologies (4) and scale$\rightleftharpoons$position mappings (5), are the sought HPC for tilt angles varying from 0° to 50° in surveillance.

4. Learning Human Perspective Context

To obtain human perspective context (HPC) for a scene, we will estimate model (4) or (5) for the average height of standing humans from observed human individuals in the scene. However, we may not be able to obtain a good model directly from the sample data since the measurements are

inaccurate and uncertain due to: (a) the variations of individuals heights, (b) the variations of body poses, (c) the errors in human shape segmentation and recognition, and (d) the spatial unbalance of samples from image space. In this paper, we propose a novel and robust method to estimate HPC from a set of inaccurate sample data.

4.1. ME-DT Algorithm

The basic idea behind the ME-DT algorithm is to employ data refinement into the process of model estimation. The proposed approach is an iterative process comprising two steps in each iteration: model estimation (ME) and data tuning (DT). The first step generates a linear HPC model from the updated sample data by using the least-square estimation (LSE), while the second step predicts and updates the sample data according to the obtained model. When the iterative process converges, we obtain the HPC model from the predicted virtual ideal sample data. In the following, the general description of the ME-DT algorithm is presented first and then the implementations for different categories of camera tilt angles are described.

Let (μ, ν) be a pair of observed data. For the foot$\rightleftharpoons$head homologies, (μ, ν) is $(\mathbf{x}_f, \mathbf{x}_h)$, and for the scale$\rightleftharpoons$position mappings, (μ, ν) is (y_b, h). Furthermore, let the pair of the mappings between the observed data be $\mathcal{H}_1$ and $\mathcal{H}_2$, *i.e.* $\nu = \mathcal{H}_1\mu$ and $\mu = \mathcal{H}_2\nu$. For foot$\rightleftharpoons$head homologies, there are $\mathcal{H}_1 = H_{hf}$ and $\mathcal{H}_2 = H_{fh}$, and for scale$\rightleftharpoons$position mappings, there are $\mathcal{H}_1 = \mathbf{m}_{sf}$ and $\mathcal{H}_2 = \mathbf{m}_{fs}$. Suppose the sample dataset is $\{\mu_i, \nu_i\}, i = 1, \cdots, N$. For initialization, let the dataset be $\{\mu_i^j, \nu_i^j\}_{i=1}^N$ with $j = 0$ (*i.e.* the first iteration), and a weight w_0 to control the data tuning. Then the following two steps can be performed repeatedly:
(a). Model Estimation (ME): The linear transformations $\hat{\mathcal{H}}_1^j$ and $\hat{\mathcal{H}}_2^j$ are generated from the dataset $\{\mu_i^j, \nu_i^j\}_{i=1}^N$ by using the least-square estimation (LSE).
(b). Data Tuning (DT): From the new models of transformation, the ideal data can be predicted as $\hat{\nu}_i^j = \hat{\mathcal{H}}_1^j\mu_i^j$ and $\hat{\mu}_i^j = \hat{\mathcal{H}}_2\nu_i^j$. Combining the original and predicted data, the virtual ideal positions are estimated as

$$\begin{aligned} \mu_i^{j+1} &= (1 - w_j)\hat{\mu}_i^j + w_j\mu_i^0 \\ \nu_i^{j+1} &= (1 - w_j)\hat{\nu}_i^j + w_j\nu_i^0 \end{aligned} \tag{6}$$

In Equ. (6), the original observation (μ_i^0, ν_i^0) is used as a dock to prevent the data moving too far away from the original position when the transformations have not become stable. The difference of the estimated data between the latest two iterations can be defined as

$$D_j = \frac{1}{N}\sum_i \left[(\hat{\mu}_i^j - \hat{\mu}_i^{j-1})^2 + (\hat{\nu}_i^j - \hat{\nu}_i^{j-1})^2\right] \tag{7}$$

This difference measure will drop significantly in a few iterations. Once it becomes stable, the best estimation should have been reached. Here, the condition for the termination is defined as $D_{j+1} \geq D_j$ or $j > 5$. If the difference is still dropping, the weight is updated as $w_{j+1} = \gamma w_j$ with $\gamma < 1$, and j is set as $j + 1$ for the next iteration. In this work, w_0 and γ are set as 0.5 and 0.7, respectively. When ME-DT terminates, the refined samples would be close to the virtual ideal positions (*i.e.* $\{\mu_i^*, \nu_i^*\} \approx \{\mu_i^j, \nu_i^j\}$). The linear transformations obtained from the virtual ideal data become the perspective model for humans in the scene (*i.e.* $\mathcal{H}_1^* \approx \hat{\mathcal{H}}_1^j$ and $\mathcal{H}_2^* \approx \hat{\mathcal{H}}_2^j$).

The performance of the ME-DT algorithm on synthetic data with various levels of noise has been evaluated. Let one ideal linear mapping be $\mu = k_\nu^*\nu + b_\nu^*$ and $\nu = k_\mu^*\mu + b_\mu^*$ with $k_\nu^* = k_\mu^* = -1$ and $b_\nu^* = b_\mu^* = 100$. In each test, 100 samples of $\{\mu_i, \nu_i\}$ are randomly generated from even distribution within [0,100] and corrupted by additive Gaussian $\mathcal{N}(0, \sigma)$, where the variance σ represents the noise level. Let the estimated parameters be $[\hat{k}_\nu, \hat{b}_\nu]$ and $[\hat{k}_\mu, \hat{b}_\mu]$, the normalized error of the estimation can be computed as $\varepsilon = \frac{1}{4}\sum_{m,n}\frac{|\hat{a}_{mn} - a_{mn}^*|}{a_{mn}^*}$ with $[a_{11}, a_{12}, a_{21}, a_{22}] = [k_\nu, b_\nu, k_\mu, b_\mu]$. At each noise level, the test is repeated 100 times randomly. The average of the normalized estimation errors with respect to the corresponding noise level can be obtained as $\bar{\varepsilon}(\sigma)$. The plot of $\bar{\varepsilon}(\sigma)$ with respect to noise levels from 5 to 40 is shown as the red curve in the left picture of Figure 3. For comparison, the average errors by the standard least-square fitting, *i.e.* the first step of the ME-DT algorithm, as well as the errors by removing 10% outliers according to the first estimation as used in [8, 7] are also plotted in blue and green colors. It can be seen that the ME-DT algorithm is much more robust to the noisy training data. To show the convergence of the ME-DT algorithm with respect to excessive noise, the example results with $\sigma = 30$ are displayed in Figure 3, where the 5 pictures from the 2nd to the right are the results of the 5 iterative steps of ME-DT algorithm. In the pictures, the red '+' marks a training data, the green line is the ideal model, the blue and black lines are the estimated mappings. The plots indicate that, by tuning the inaccurate data according to the previous estimation, ME-DT can converge to the ideal model quickly even on data with excessive noise. The most significant difference of the ME-DT to existing robust estimation methods (e.g. ME, RANSAC, and LMS) is that it tunes the training data to guess the ideal data from the inaccurate data.

4.2. Estimate HPC Using ME-DT Algorithm

For a scene observed from a *large* camera tilt angle, the head and foot positions of human samples can be directly used to estimate the foot$\rightleftharpoons$head homologies. Let $\mathbf{x}_h = (x_h, y_h)$ and $\mathbf{x}_f = (x_f, y_f)$ be the head and foot points of a sample located along the principal axis of the silhouette, and $H_{hf} = [a_{mn}]_{3\times 3}$ be the foot$\rightarrow$head homol-

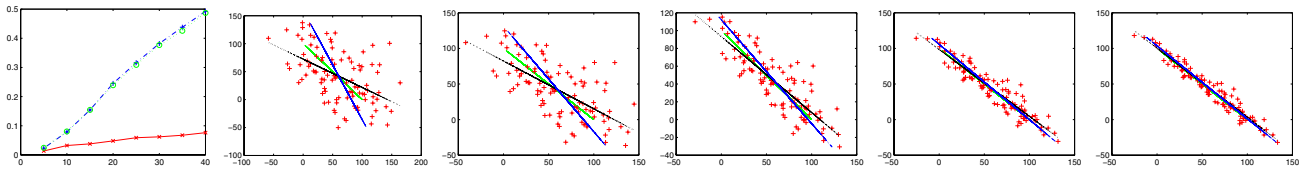

Figure 3. The evaluation of the ME-DT algorithm on synthetic data.

ogy. Since H_{hf} is a mapping between homogeneous coordinates, we can assume $a_{33} = 1$. Then, the left equation in (4) can be rewritten as

$$\begin{bmatrix} x_f & y_f & 1 & 0 & 0 & 0 & -x_f x_h & -y_f x_h \\ 0 & 0 & 0 & x_f & y_f & 1 & -x_f y_h & -y_f y_h \end{bmatrix} \mathbf{a} = \begin{bmatrix} x_h \\ y_h \end{bmatrix} \tag{8}$$

where $\mathbf{a} = [a_{11}\ a_{12}\ a_{13}\ a_{21}\ a_{22}\ a_{23}\ a_{31}\ a_{32}]^\tau$ are the rest elements of H_{hf}. From (8), the least-square estimation $\hat{H}_{hf}$ can be obtained from a set of N samples $\{\mathbf{x}_{hi}, \mathbf{x}_{fi}\}_{i=1}^N$. Similarly, $\hat{H}_{fh}$ can be obtained. Using such least-square estimations in the Model Estimation step in the ME-DT algorithm, the robust estimation of the foot⇌head homologies can be obtained.

For a scene observed from a *small* camera tilt angle, the foot⇌head homologies are still used to describe the human perspective context, but there are two differences in the implementation of ME-DT algorithm to estimate the homologies. First, the x-position of the center is used for both the foot and the head, *i.e.* x_f=x_h=x_c. This is because for a small tilt angle, the principal axis of an upright human anywhere in an image is close to a vertical line. The biases of the head's and foot's x-positions to the center point are caused by the variations of poses and segmentation errors. They often causes the failure in homology estimation. Second, only the foot→head homology estimate $\hat{H}^j_{hf}$ is generated from the sample data using the least-square fitting. Since in vertical direction, the head→foot homology is a linear mapping from a narrow range to a wide range. It is easy to lead to large divergence from the true model when noisy and inaccurate sample data is used. The head→foot homology is obtained as $\hat{H}^j_{fh} = (\hat{H}^j_{hf})^{-1}$.

In the case of *zero* camera tilt angle, the scale⇌position mappings in (5) are used to describe HPC. The sample data is a pair of foot position and the height of the body, *i.e.* $(\mu_i, \nu_i) = (y_{bi}, h_i)$. In Model Estimation step, the standard least-square fitting is used to estimate $\hat{\mathbf{m}}^j_{fs}$ and $\hat{\mathbf{m}}^j_{sf}$.

5. Efficient Human Detection

The aim of visual surveillance is to interpret human related events in the scene. Human detection plays an important role to achieve such a goal. Given an image $I(\mathbf{x})$, detection of a human standing at $\mathbf{x}_f$ is to compute the likelihood $P(I, \mathbf{x}_f|A)$, where A denotes a human appearance model. Without any prior knowledge of the scene, the likelihood should be evaluated over all possible **s**cales, **o**rientations, and viewing **a**ngles, that is

$$P(I, \mathbf{x}_f|A) = \arg \max_{\{a,s,o\}} P(I, \mathbf{x}_f|A(a, s, o)) \tag{9}$$

With learned HPC, the scale, orientation, and appearance with respect to the viewing angle can be determined as $A_a(s_{\mathbf{x}_f}, o_{\mathbf{x}_f})$. Hence, there is

$$P(I, \mathbf{x}_f|A) \propto P(I|A_a(s_{\mathbf{x}_f}, o_{\mathbf{x}_f}), \mathbf{x}_f) \tag{10}$$

This means, with learned HPC, there is no need to compute the likelihood over multiple scales, orientations, and appearances at each position. This will improve not only the efficiency but also the accuracy for human detection.

In this paper, the HOG human detector [3] is employed. First, two HOG human detectors for *large* and *small* camera tilt angles are trained offline, *i.e.* $A_a(\cdot,\cdot), a = 1, 2$. For a specific scene, once the camera tilt angle is recognized online, the corresponding HOG detector will be applied. To adapt to great scale variations of humans in some scenes, the detection window is divided into 4×8 cells whereas the size of the cells varies from 4×4 to 20×20 pixels. Therefore, we have detection windows of 17 scale levels from 16×32 to 80×160 pixels. Each detection window consists of 3×7 blocks formed by 2×2 cells in a sliding fashion.

Using learned HPC, a scene-adaptive grid on the floor which determines the scale and orientation of a human in a position in the image can be generated as shown in Figure 4. The grid is generated row by row from the bottom of the image. The horizontal distance between two adjacent grid points in the same row is the half of human width. The vertical distance between two adjacent rows is set so that there is just $\beta\%$ (*e.g.* 85%) vertical overlap of human heights in the two rows. Since the grid density is determined by the related human scales, the number of the grid points, or detection windows, is not much related to the image size.

For a new incoming image, Human detection is performed one-by-one at each grid point. Let a grid point be the foot position of a possible person and denoted as $\mathbf{x}_f$. The corresponding head position $\mathbf{x}_h$ can then be estimated using the learned HPC. The height $\|\mathbf{x}_f - \mathbf{x}_h\|$ determines the scale of the detection window. The orientation of the window is $\theta = \tan^{-1}(\frac{x_h - x_f}{y_h - y_f})$. The sub-region within the

Method	*large*	*small*	*zero*	average
1 model	14.9%	343.2%	144.9%	105.6%
2 models	14.9%	18.7%	8.8%	14.4%
ME-DT	8.8%	5.2%	5.4%	7.4%

Table 2. The evaluation results on HPC learning.

window is rotated to the upright position and the HOG feature vector from the window can be obtained. The feature vector is then fed to the SVM model for classification.

6. Experiments and Evaluations

The proposed method has been tested and evaluated on more than 15 scenes from well-known banchmark datasets and real CCTV systems. The efficient human detection consists of two stages: automatic HPC learning and HPC-constrained human detection. The performance of these two parts are evaluated separately in the following.

The performance of HPC learning

Once a camera is set up, a robust method of adaptive background subtraction is applied. Foreground regions are extracted and shadows are suppressed. Samples of moving persons are selected according to the human shape model described in subsection 2.1. To evaluated the robustness for learning HPC, three approaches are compared: (1) single model (foot⇌head homology) and LSE; (2) multiple models (foot⇌head homology and scale⇌position mapping) and LSE; (3) multiple model and ME-DT. The error rate of the estimation is defined as

$$ER = \frac{1}{N}\sum_{i=1}^{N}\frac{|\mathbf{x}_{fi}-\hat{\mathbf{x}}_{fi}|+|\mathbf{x}_{hi}-\hat{\mathbf{x}}_{hi}|}{2|\mathbf{x}_{fi}-\mathbf{x}_{hi}|} \qquad (11)$$

where $(\mathbf{x}_{fi}, \mathbf{x}_{hi})$ is the ith sample and $(\hat{\mathbf{x}}_{fi}, \hat{\mathbf{x}}_{hi})$ is the corresponding estimation. The average error rates of 3 approaches on the 15 scenes are listed in Table 2. The performance of ME-DT is the best. The error rate is 3.6% when manually annotated inputs are used for test. The error rate of 3.6% is caused by variations of human heights.

The performance of human detection

One benefit of applying HPC to human detection is the great reduction of detection windows for each image. With no knowledge of human appearance in the scene, we have to scan the image with multi-scale detection windows, *e.g.*, for an image of 320×240 pixels, over 200,000 windows of 5 scales are used [13, 14]. Applying HPC, less than 2,000 windows are enough for 17 scale levels and orientation variations. Three examples of scene-adaptive grids for the 3 categories of camera tilt angles are shown in Figure 4, where the red dots represents foot positions and green dots are

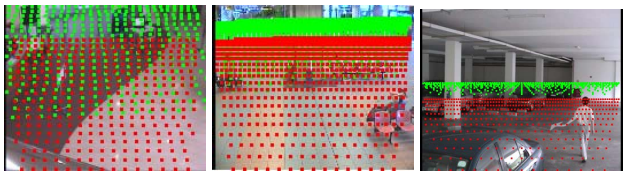

Figure 4. Examples of scene-adaptive grids for *large*, *small*, and *zero* camera tilt angles.

the corresponding head positions. A pair of red and green dots denotes a detection window. The average and maximum numbers of detection windows from the 15 scenes are 810 and 1826. This means, compared to the conventional method, less than 1% detection windows is required.

Another benefit of applying HPC is the improvement on human detection. This can be seen from the comparison with conventional method on some examples displayed in Figure 5, where the postprocessing of clustering is not performed. Firstly, human like objects but much larger, smaller or taller, such as the trees and the pillars in the first three examples, would not be detected. Secondly, when the camera tilt angle is large, the leaning of body orientation for humans on both sides of the scene is significant. Extracting HOG features from the rotated detection window according to HPC can improve the human detection performance, as shown in the 2nd column. Thirdly, since the proposed method does not depend on estimation of vanishing horizon, it can also be used for the scenes where the ground surface is not parallel to the horizontal sea plane. One example is shown in the 3rd column. Fourthly, since HPC helps avoiding some impossible false positives, we can use a low threshold for human detection. In this case, the partial occluded humans can be detected with a HOG detector trained with no occlusion samples. As shown in the last two columns, the persons partially occluded by car door or other person can be detected. In addition, human detection under the guide of HPC also provides clues about 3D spatial correlations of humans in the group.

The systematic evaluation of HPC-constrained human detection is also performed. We sampled the detection results every 50 frames evenly from the 15 test sequences. In the evaluation, each person of more than 30% being visible is counted as a valid person. Persons of too small scales (height$<$20 pixels) are not counted. For a valid person, if the overlap of the body and a detected window is over 50%, it is accepted as detected. If less than 30% coverage of a detected window is related to a valid person, it is labelled as a false positive. Cluster of overlapped detections over the same background object is counted as one false positive since no postprocessing is performed. The final results of the statistics and sampled images from the 15 scenes can be found in the supplementary document. On average,

Figure 5. The examples of human detection with and without HPC constraints int the lower and upper rows.

our method achieved **91.2%** detection rate with **2.05** false positives per frame (FPPF). Existing methods can achieve 80-90% detection rate at 10^{-4} false positives per window (FPPW) which corresponds to 2-3 false positives per image [13, 14, 3, 17]. However, our result is obtained with HOG detectors trained by less than 200 samples (positive and negative samples together) and we counted small and partially occluded persons in the evaluation. The majority of false positives are detected with small windows for humans of heights between 20 to 40 pixels, which are too small to be detected by existing methods.

7. Conclusions

In this paper, a novel and practical way for unsupervised learning of human perspective context (HPC) for a scene is proposed. It contains three parts: estimation of camera tilt angle, modeling of HPC for camera tilt angles ranging from 0° to 50°, and the ME-DT algorithm for robust estimation of HPC from automatically selected human samples. An efficient approach for human detection constrained by HPC is also proposed. Experiment results show the robustness of HPC learning and the benefits of applying HPC for human detection on videos captured with large range of camera tilt angles. If the camera tilt angle or focus length has been changed, the difference between the detected and estimated head and foot positions will increase greatly. In this case, the system can start to learn the HPC again. This means the method is suitable for autonomous intelligent video surveillance systems. In future work, we aim to apply the learned HPC to improve scene analysis, foreground refinement, shadow suppression, object classification, tracking, and identification in automated video surveillance.

References

[1] B. Caprile and V. Torre. Using vanishing points for camera calibration. *IJCV*, 4:127–140, 1990. 1

[2] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. *IJCV*, 40:123–148, 2000. 1, 4

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR*, 1:886–893, 2005. 6, 8

[4] A. Elgammal and L. Davis. Probabilistic framework for segmenting people under occlusion. *ICCV*, 2:145–152, 2001. 3

[5] O. Faugeras. *Three Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, 1993. 1

[6] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. *CVPR*, 2:2137–2144, 2006. 1

[7] I. Junejo and H. Foroosh. Robust auto-calibration from pedestrians. *AVSS*, 1:92–97, 2006. 1, 2, 5

[8] N. Krahnstoever and P. Mendonca. Bayesian autocalibration for surveillance. *ICCV*, 2:1858–1865, 2005. 1, 2, 5

[9] L. Li, I. Gu, M. Leung, and Q. Tian. Adaptive background subtraction based on feedback from fuzzy classification. *Optical Engineering*, 43:2381–2394, 2004. 3

[10] F. Lv, T. Zhao, and R. Nevatia. Camera calibration from video of a walking human. *IEEE T-PAMI*, 28:1513–1518, 2006. 1, 2

[11] J. Semple and G. Kneebone. *Algebraic Peojection Geometry*. Oxford Classic Texts in the Physical Sciences. Clarendon Press, Oxford, UK, 1998, Originally Published in 1952. 4

[12] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis, and Machine Vision*. PWS Publishing, 1999. 3

[13] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *IJCV*, 63:153–161, 2005. 7, 8

[14] Y. Wu and T. Yu. A field model for human detection and tracking. *IEEE T-PAMI*, 28:753–765, 2006. 7, 8

[15] Z. Zhang. A flexible new technique for camera calibration. *IEEE T-PAMI*, 22:1330–1334, 2000. 1

[16] T. Zhao, R. Nevatia, and F. Lv. Segmentation and tracking of multiple humans in complex situations. *CVPR*, 2:194–201, 2001. 1

[17] Q. Zhu, S. Avidan, M.-C. Yeh, and K.-T. Cheng. Fast human detection using a cascade of histograms of oriented gradients. *CVPR*, 2:1491–1498, 2006. 8