

METHOD AND SYSTEM FOR LEARNING HPC FOR HUMAN DETECTION IN VIDEO DATA

5 FIELD OF INVENTION

The present invention relates broadly to methods and systems of learning human perspective content (HPC) for human detection in video data, and to a data storage medium having computer code means for instructing a computing device to
10 execute a method of learning HPC for human detection in video data.

BACKGROUND

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

In addition, one aim of video surveillance is to interpret human-related events
25 and activities in the videos. Therefore, human detection plays a very important role to achieve such goal. Given an input image $I(\mathbf{x})$, detecting a human object standing at position $\mathbf{x}_f$ in the image can be expressed as evaluating a probability $P(I, \mathbf{x}_f | h)$, where h denotes a human appearance model. According to the perspective projection transformation, the appearance of a human object in an image is determined by three
30 parameters: the tilt angle of the camera to the scene, the scale (i.e. size), and the principal orientation of the human at the position in the image. Without any prior knowledge of the scene, the probability should be evaluated over all the possible values of the three parameters, that is

$$P(I, \mathbf{x}_f | h) = \arg \max_{\{a, s, o\}} P(I, \mathbf{x} | h(a, s, o)) \quad (1)$$

where a , s , and o represent the tilt angle of the camera, the scale and the orientation of the human object respectively.

5 As appreciated by a skilled person in the art, the probability P in Equation (1) cannot be computed directly because a very large search space is needed for optimisation and many false positives may be generated due to the lack of perspective constraint on the specific view of the human object. To make the problem more manageable, assumptions are usually made. In most prior art approaches, it is assumed
10 that the scene is observed from an angle that is nearly horizontal, which means the tilt angle of the camera with respect to the ground plane is small (between 0° to 20°). In such cases, the variations of human appearances caused by view angles can be ignored, and the orientations of human objects are almost all upright in the images. Even then, a huge number (at least a hundred thousands) of multi-scale detection windows
15 are usually required to scan the whole image.

In principle, the perspective projection about a scene is precisely described by the camera parameters and three-dimensional (3D) world geometry of the scene. The conventional approaches to establish the camera parameters are known as camera
20 calibration. The conventional methods for camera calibration assume that a few special calibration objects, e.g. long parallel lines, or measurements of enough 3D points in the scene are available. Using the conventional methods, it is possible to obtain accurate estimates of the camera parameters. However, such manual techniques are not adequate for automated systems, e.g. automated video surveillance systems with many
25 widely deployed cameras.

Recently, some approaches with varying degrees of automation have been proposed based on pedestrian data and on the understanding that the vanishing points of parallel lines in the scene can be used to calculate camera parameters. Table 1
30 provides a summary of some of these prior art approaches. Prior Art A uses a tracking approach to select samples of a person walking along two different directions. 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. In Prior Art B, two harmonic homologies are introduced to derive linear equations for the focal length of the camera since both horizontal and vertical vanishing points can be obtained from a pair of images of the tracked person. Since 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 around in the scene. Therefore, they can only be considered as semi-automatic methods.

Table 1 (Prior Art Comparison)

Method	Learning	Assumption	Angle	Estimation	Measurement	Tracking	Robustness
Prior Art A	semi-auto	yes	5°-30°	non-linear	real	yes	weak
Prior Art B	semi-auto	yes	10°-25°	linear	real	yes	strong
Prior Art C	auto	yes	8°-25°	non-linear	real	no	strong

Prior Art C proposes a fully automatic self-calibration method from detected humans in which foot-to-head homology is first estimated and then decomposed to extract the vanishing point and vanishing horizon for calibration. The camera parameters are obtained from a maximum a posteriori (MAP) estimation under a Bayesian framework. Metropolis sampling for random optimization is used to solve the MAP problem. However, prior knowledge about camera parameters and sample distributions are needed. In addition, stochastic optimization is robust to noisy data, but it requires not only intensive computations but also good initial estimates.

In the prior art approaches described above, while camera parameters are usually used in the formal description of perspective projection, it is difficult to estimate the parameters 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. Furthermore, the ranges for the camera angle are rather limited.

Therefore, a need exists to provide a method and system that seek to address at least one of the above problems.

SUMMARY

In accordance with a first aspect of the present invention, there is provided a method of learning human perspective context (HPC) for human detection in video data, the method comprising the steps of:

selecting a plurality of samples based on human detection from the video data;

categorizing a camera tilt angle into one of two or more categories based on the selected samples; and

learning the HPC from the selected samples;

wherein the samples are filtered based on the categorization of the camera tilt angle prior to learning the HPC, a model calculation for the HPC is based on the categorization of the camera tilt angle, or both.

The selecting of the samples may comprise:

detecting foreground objects in the video data;

determining probabilities of respective ones of the foreground objects corresponding to one of at least two shape feature models; and

selecting respective objects as one of the samples if either probability meets a first threshold.

The determining of the probabilities of the respective ones of the foreground objects belonging to one of the shape feature models may be based on a shape feature vector of the foreground object.

The categorizing of the camera tilt angle may comprise determining head and foot positions of the respective selected samples, and determining likelihoods for the plurality of selected samples belonging into respective ones of two or more classes of head and foot positions corresponding to the respective categories of the tilt angle.

The categorizing of the camera tilt angle may further comprise determining said one category based on both the likelihoods for the plurality of selected samples belonging into respective ones of the two or more classes of head and foot positions

and the probabilities of respective ones of the foreground objects corresponding to one of at the least two shape feature models.

5 The filtering of the samples may comprise filtering by one of the shape feature models corresponding to the categorization of the camera tilt angle.

The model calculation for the HPC may use either foot $\Leftrightarrow$ head homologues or scale $\Leftrightarrow$ position mapping to describe the HPC depending on the categorisation of the camera tilt angle.

10

The model calculation may further use different implementations of describing the HPC using the foot $\Leftrightarrow$ head homologues depending on the categorisation of the camera tilt angle.

15

The model calculation for the HPC, either foot $\Leftrightarrow$ head homologues or scale $\Leftrightarrow$ position mapping, may comprise using a model estimation - data tuning (ME-DT) algorithm.

20

The method may further comprise applying the learnt HPC in human detection from subsequent video data.

25

The applying of the learnt HPC may comprise generating a scene-adaptive grid which defines a scale and orientation of a human in a position in an image, and limiting detection windows of a human detection algorithm based on the scene-adaptive grid.

The human detection algorithm may be selected from two or more algorithms based on the categorization of the camera tilt angle.

30

In accordance with a second aspect of the present invention, there is provided a system for learning human perspective context (HPC) for human detection in video data, the system comprising:

means for selecting a plurality of samples based on human detection from the video data;

means for categorizing a camera tilt angle into one of two or more categories based on the selected samples; and

means learning the HPC from the selected samples;

5 wherein the means for learning the HPC filtered the sample based on the categorization of the camera tilt angle prior to learning the HPC, uses a model calculation for the HPC based on the categorization of the camera tilt angle, or both.

The means for selecting of the samples may:

detect foreground objects in the video data;

10 determine probabilities of respective ones of the foreground objects corresponding to one of at least two shape feature models; and

select respective objects as one of the samples if either probability meets a first threshold.

15 The determining of the probabilities of the respective ones of the foreground objects belonging to one of the shape feature models may be based on a shape feature vector of the foreground object.

20 The means for categorizing the camera tilt angle may determine head and foot positions of the respective selected samples, and may determine likelihoods for the plurality of selected samples belonging into respective ones of two or more classes of head and foot positions corresponding to the respective categories of the tilt angle.

25 The means for categorizing the camera tilt angle may further determine said one category based on both the likelihoods for the plurality of selected samples belonging into respective ones of the two or more classes of head and foot positions and the probabilities of respective ones of the foreground objects corresponding to one of at the least two shape feature models.

30

The means for learning the HPC may filter the samples by one of the shape feature models corresponding to the categorization of the camera tilt angle.

The means for learning the HPC may use either foot $\Leftrightarrow$ head homologies or scale $\Leftrightarrow$ position mapping to describe the HPC depending on the categorisation of the camera tilt angle.

- 5 The means for learning the HPC may use different implementations of describing the HPC using the foot $\Leftrightarrow$ head homologies depending on the categorisation of the camera tilt angle.

- 10 The system may further comprise means for applying the learnt HPC in human detection from subsequent video data.

- 15 The means for applying the learnt HPC may generate a scene-adaptive grid which defines a scale and orientation of a human in a position in an image, and limits detection windows of a human detection algorithm based on the scene-adaptive grid.

The human detection algorithm may be selected from two or more algorithms based on the categorization of the camera tilt angle.

- 20 In accordance with a third aspect of the present invention, there is provided a data storage medium having computer code means for instructing a computing device to execute a method of learning human perspective context (HPC) for human detection in video data, the method comprising the steps of:

- 25 selecting a plurality of samples based on human detection from the video data;

categorizing a camera tilt angle into one of two or more categories based on the selected samples; and

learning the HPC from the selected samples;

- 30 wherein the samples are filtered based on the categorization of the camera tilt angle prior to learning the HPC, a model calculation for the HPC is based on the categorization of the camera tilt angle, or both.

BRIEF DESCRIPTION OF THE DRAWINGS

Embodiments of the invention will be better understood and readily apparent to one of ordinary skill in the art from the following written description, by way of example only, and in conjunction with the drawings, in which:

Figure 1 shows a flowchart illustrating a method of learning human perspective context (HPC) for human detection in video data.

Figure 2 shows perspective of human objects in a scene at various positions.

Figure 3 is a schematic diagram of a perspective scene observed from a tilt angle close to 0° for determining scale $\Leftrightarrow$ position mapping according to an example embodiment.

Figure 4 shows schematic diagrams of perspective of a scene as observed from a large angle, a small angle and a near-zero angle respectively.

Figure 5 shows shapes of human objects obtained from example images of various scenes.

Figure 6 shows membership functions for the 3 tilt angle categories with respect to the ratio of vertical distribution ranges of head and foot according to an example embodiment.

Figure 7 shows a detailed flowchart of a method to categorise a camera tilt angle for determining a perspective projection model.

Figure 8A-8B show the performance of the ME-DT algorithm on synthetic data with various noise level according to an example embodiment.

Figure 9 is a flowchart of an implementation of a perspective-constrained human detection method according to an example embodiment.

Figure 10 shows a series of images of various scenes and the corresponding grids generated thereon according to an example embodiment.

Figure 11 shows a series of human detection results using a conventional method and the corresponding results using the method according to an example embodiment.

Figure 12 is a schematic diagram of a computer for implementing the method of Figure 1 according to an example embodiment.

DETAILED DESCRIPTION

The example embodiments provide, *inter alia*, a method for learning human perspective context (HPC) for a variety of scenes monitored by surveillance cameras and applying such scene-associated context information for e.g. efficient and accurate human detection, foreground object classification, and refinement of segmentation of human objects. It is found that two linear models, i.e. foot $\Leftrightarrow$ head homologies and scale $\Leftrightarrow$ position mappings, can advantageously be used to describe human perspective context about a scene for camera tilt angles from 0° to 50° to the ground plane. When the camera tilt angle varies from large to small, the vertical range of human head positions in the image shrinks faster than that of the foot positions. The example embodiment proposes an approach to categorise the 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 corresponding to the tilt angle.

As the automatically selected samples may be uncertain and noisy due to the variations of human individual heights and poses, and because of errors of foreground segmentation and shape recognition, a model estimation - data tuning (ME-DT) algorithm is also provided to learn human perspective context from such training data. The ME-DT algorithm can not only estimate the model parameters from the training data but also tunes the data to achieve a good head-foot correlation. The algorithm is implemented for the cases of three tilt angles, where different learning strategies are

used for different cases. The ME-DT algorithm can generate good estimation of human perspective context from the data set where the conventional linear estimation fails.

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, a Histogram of Oriented Gradients (HOG) detection window of corresponding scale and orientation determined by the HPC is applied. It is observed that less than 1% of detection window is required as compared with conventional human detection methods.

Figure 1 shows a flowchart 100 illustrating a method of learning HPC for human detection in video data, according to an example embodiment. At step 102, a plurality of samples is selected from the video data based on human detection. At step 104, a camera tilt angle is categorized into one of two or more categories based on the selected samples. At step 106, the HPC is learnt from the selected samples, wherein the samples are filtered based on the categorization of the camera tilt angle prior to learning the HPC, a model calculation for the HPC is based on the categorization of the camera tilt angle, or both.

Some portions of the description which follows are explicitly or implicitly presented in terms of algorithms and functional or symbolic representations of operations on data within a computer memory. These algorithmic descriptions and functional or symbolic representations are the means used by those skilled in the data processing arts to convey most effectively the substance of their work to others skilled in the art. An algorithm is here, and generally, conceived to be a self-consistent sequence of steps leading to a desired result. The steps are those requiring physical manipulations of physical quantities, such as electrical, magnetic or optical signals capable of being stored, transferred, combined, compared, and otherwise manipulated.

Unless specifically stated otherwise, and as apparent from the following, it will be appreciated that throughout the present specification, discussions utilizing terms such as "estimating", "calculating", "determining", "performing", "generating", "initializing", "outputting", or the like, refer to the action and processes of a computer system, or similar electronic device, that manipulates and transforms data represented as physical quantities within the computer system into other data similarly represented as physical

quantities within the computer system or other information storage, transmission or display devices.

5 The present specification also discloses apparatus for performing the operations of the methods. Such apparatus may be specially constructed for the required purposes, or may comprise a general purpose computer or other device selectively activated or reconfigured by a computer program stored in the computer. The algorithms and displays presented herein are not inherently related to any particular computer or other apparatus. Various general purpose machines may be used with programs in
10 accordance with the teachings herein. Alternatively, the construction of more specialized apparatus to perform the required method steps may be appropriate. The structure of a conventional general purpose computer will appear from the description below.

15 In addition, the present specification also implicitly discloses a computer program, in that it would be apparent to the person skilled in the art that the individual steps of the method described herein may be put into effect by computer code. The computer program is not intended to be limited to any particular programming language and implementation thereof. It will be appreciated that a variety of programming languages and coding thereof may be used to implement the teachings of the disclosure
20 contained herein. Moreover, the computer program is not intended to be limited to any particular control flow. There are many other variants of the computer program, which can use different control flows without departing from the spirit or scope of the invention.

25 Furthermore, one or more of the steps of the computer program may be performed in parallel rather than sequentially. Such a computer program may be stored on any computer readable medium. The computer readable medium may include storage devices such as magnetic or optical disks, memory chips, or other storage devices suitable for interfacing with a general purpose computer. The computer readable medium may also include a hard-wired medium such as exemplified in the Internet
30 system, or wireless medium such as exemplified in the GSM mobile telephone system. The computer program when loaded and executed on such a general-purpose computer effectively results in an apparatus that implements the steps of the preferred method.

The method and system of the example embodiment can be implemented on a computer system 1200, schematically shown in Figure 12. It may be implemented as software, such as a computer program being executed within the computer system 1200, and instructing the computer system 1200 to conduct the method of the example embodiment.

The computer system 1200 comprises a computer module 1202, input modules such as a keyboard 1204 and mouse 1206 and a plurality of output devices such as a display 1208, and printer 1210.

The computer module 1202 is connected to a computer network 1212 via a suitable transceiver device 1214, to enable access to e.g. the Internet or other network systems such as Local Area Network (LAN) or Wide Area Network (WAN).

The computer module 1202 in the example includes a processor 1218, a Random Access Memory (RAM) 1220 and a Read Only Memory (ROM) 1222. The computer module 1202 also includes a number of Input/Output (I/O) interfaces, for example I/O interface 1224 to the display 1208, and I/O interface 1226 to the keyboard 1204.

The components of the computer module 1202 typically communicate via an interconnected bus 1228 and in a manner known to the person skilled in the relevant art.

The application program is typically supplied to the user of the computer system 1200 encoded on a data storage medium such as a CD-ROM or flash memory carrier and read utilising a corresponding data storage medium drive of a data storage device 1230. The application program is read and controlled in its execution by the processor 1218. Intermediate storage of program data maybe accomplished using RAM 1220.

In a typical scene, humans usually stand and walk on the ground surface. Let h_a represent the human appearance observed from a tilt angle a to the scene. Equation (1) can be written as

$$P(I, \mathbf{x}_f | h) = \arg \max_{\{s, o\}} P(I, \mathbf{x} | h_a(s, o)) \quad (2)$$

Additionally, according to Bayesian law,

$$P(I, \mathbf{x}_f | h_a(s, o)) \propto P(I | h_a(s, o), \mathbf{x}_f) P(h_a(s, o) | \mathbf{x}_f) \quad (3)$$

Once the camera is fixed, the perspective projection transformation for the scene can be determined. The position, scale and orientation of a human object in an image are determined by where he stands on the ground surface according to the perspective projection. Let $\mathbf{x}_f$ be the foot position of a human object in an image. The scale and orientation of the person in the image can be determined as $o_{\mathbf{x}_f}$ and $s_{\mathbf{x}_f}$ according to human metrics and the perspective projection transformation for the scene. Then, it can be assumed that

$$P(h_a(s_x, o_x) | \mathbf{x}_f) = \begin{cases} 1, & \text{if } \mathbf{x} = \mathbf{x}_f \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Combining Equations (1) and (4), the probability of observing a human standing at $\mathbf{x}_f$ in an image becomes

$$P(I, \mathbf{x}_f | h) \propto P(I | h_a(s_{\mathbf{x}_f}, o_{\mathbf{x}_f}), \mathbf{x}_f) \quad (5)$$

Figure 2 shows perspective of human objects 202 in a scene 200 at various positions. It can be seen that, at any given tilt angle, the size and orientation of the human objects 202 change depending on the respective positions in the scene. From Equation 5, once the perspective of human objects 202 in the scene 200 is obtained, it may be possible to detect human object at each position in the image with a human appearance model of known scale and orientation for a specific camera tilt angle to the scene.

The perspective of human objects in a scene (i.e. Equation (4)) also provides a spatial constraint of human objects in the scene. It provides a way to transform the

image measurement of a foreground object to its physical measurement with respect to human metrics. Hence, it can provide very important context information to verify foreground segmentation results, classify foreground objects, and understand human-related activities in the scene.

5

Human perspective context (HPC) is the knowledge of perspective projection of human in the scene. In the example embodiment, two prospective projection models are provided. The perspective projection determines the transformation from a world coordinate system to the image coordinate frame. Using homogeneous coordinates, let $\tilde{\mathbf{X}} = [X, Y, Z, 1]^T$ be a point in a real world coordinate system. The projection transformation is described as $\tilde{\mathbf{x}} = H\tilde{\mathbf{X}}$ where $\tilde{\mathbf{x}} = [\tilde{x}, \tilde{y}, \tilde{z}]^T$. The image position is obtained as $x = \tilde{x}/\tilde{z}$ and $y = \tilde{y}/\tilde{z}$.

Let the XY plane be the ground surface. The foot position of a person can be denoted as $\mathbf{X}_f = [X, Y, 0]^T$. Since $Z = 0$, the 3rd column in H may be removed. The perspective projection becomes $\tilde{\mathbf{x}}_f = H_f \tilde{\mathbf{X}}_f$ with $\tilde{\mathbf{X}}_f = [X, Y, 1]^T$. Similarly, if the XY plane is translated upwards to the plane level with the head, then $\tilde{\mathbf{x}}_h = H_h \tilde{\mathbf{X}}_h$ where both H_f and H_h are 3×3 matrices. It should be noted that for a standing person, the foot location on the ground plane corresponds to exactly one location in the head plane such that $\tilde{\mathbf{X}}_f = \tilde{\mathbf{X}}_h$. The transformation can be expressed as

$$\tilde{\mathbf{x}}_h = H_{hf} \tilde{\mathbf{x}}_f \text{ and } \tilde{\mathbf{x}}_f = H_{fh} \tilde{\mathbf{x}}_h \quad (6)$$

where $H_{hf} = H_h H_f^{-1}$ and $H_{fh} = H_f H_h^{-1}$ are 3×3 matrices and $H_{hf} H_{fh} = H_{fh} H_{hf} = I$.

As will be appreciated by a person skilled in the art, the linear mapping of H_{hf} or H_{fh} is called the homology between two parallel planes in real world. The homology can also be derived from e.g. the plane vanishing horizon, the vertical vanishing point and the physical distance between the two planes in the scene.

30

Further, it will be appreciated that foot $\Leftrightarrow$ head homologies exist except when the camera is placed at the same height as the head plane such that the head plane is projected into the vanishing horizon and H_h becomes singular. Since both H_m and H_h involve H_h , the foot $\Leftrightarrow$ head homologies become invalid. In some indoor surveillance systems, the camera position may be limited by ceiling. In addition, to achieve a large depth coverage, the camera may be installed almost horizontally, i.e. the tilt angle is close to zero. In such cases, the foot $\Leftrightarrow$ head homologies may not be suitable.

In the example embodiment, a simple linear model of scale $\Leftrightarrow$ position mapping is used for small tilt angles. As shown in Figure 3, when the camera tilt angle is close to 0° (e.g. below about 10°), the upright humans are nearly parallel to the image plane. Let y_0 be the vertical location of the vanishing horizon in the image, 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 respectively. The corresponding head positions are y_h^1 and y_h^2 . According to the relations for similar triangles $\Delta y_0 y_f^1 y_h^1$ and $\Delta y_0 y_f^2 y_h^2$, it is possible to 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 $\Leftrightarrow$ position mappings can be expressed as

$$s = \mathbf{m}_s \tilde{\mathbf{y}}_f \text{ and } y_f = \mathbf{m}_{fs} \tilde{\mathbf{s}} \quad (7)$$

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

In the example embodiment, the two models, i.e. foot $\Leftrightarrow$ head homologies as defined in Equation (6) and scale $\Leftrightarrow$ position mapping as defined in Equation (7), are used to describe human perspective context for tilt angles varying from 0° to 50° in surveillance.

Figure 4 shows schematic diagrams of perspective of a scene as observed from (a) a large angle, (b) a small angle and (c) a near-zero angle respectively. It can be seen that when observed from a large tilt angle (Figure 4a), 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 positions in the scene. Moreover, the body orientations near the left and right edges of the image differ due to the effect of perspective. On the other hand, when the tilt angle is near zero (Figure 4c), the vertical range of head positions in the image is much less than that of the corresponding foot positions while variations in body orientation due to the effect of perspective become marginal. Hence, it has been recognised by the inventors that different models and strategies are advantageous for determining HPC at different tilt angles.

Based on that recognition, in the example embodiment, the camera tilt angles are classified into three categories, i.e. large, small, and near-zero angles. The ranges of the angles for the categories are about $25^\circ - 50^\circ$, $8^\circ - 25^\circ$ and $0^\circ - 8^\circ$ respectively. It should be appreciated that a different number $n \geq 2$ of categories and different angle ranges may be selected.

The camera tilt angle can be estimated based on visual clues e.g. silhouettes (i.e. shapes) of human samples and vertical distributions of head and foot positions of human samples in the image.

When observed from different camera tilt angles, the silhouettes of a standing or 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 variations of view angles to the body (e.g. the front view or the side view). To estimate the camera tilt angle from the silhouettes of observed human samples, the shape feature is preferably able to distinguish the 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.

In the example embodiment, three shape features are employed on the silhouettes of human samples obtained from e.g. background subtraction. Firstly, a moment-based shape descriptor captures the shape features of different poses of standing and walking persons. A 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, a aspect ratio of an observed human shape can reflect the camera tilt angle. By integrating the aspect ratio, head - shoulder width ratio, and moments into a shape feature clue, single standing or walking persons in the scene can be recognised and the correlation between the camera tilt angle and observed human shapes can be characterised.

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

$$\mu_{jk} = \sum_{(x,y)} (x-\bar{x})^j (y-\bar{y})^k R(x,y) \quad (8)$$

where $[\bar{x}, \bar{y}]^T$ is the gravity centre of the shape. Using equation (8) the normalised 2nd-order central moments are defined as

$$\phi_1 = \frac{\sqrt{|\mu_{11}|}}{|R|}, \quad \phi_2 = \frac{\sqrt{|\mu_{20}|}}{|R|}, \quad \phi_3 = \frac{\sqrt{|\mu_{02}|}}{|R|} \quad (9)$$

where $|R|$ is the size of the shape.

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, the left and right ends x_l and x_r can be obtained, and from the projection on the vertical axis, the top and bottom ends y_t and y_b can be obtained. Preferably, the tails on both sides of a histogram of x or y values are cut at 10% of the peak height to be robust to noise and shape details. The aspect ratio of the two-dimensional (2D) shape is computed as $r_a = (x_r - x_l) / (y_t - y_b)$. For a standing person, the neck position can be assumed to be at 81% of the height from the foot. 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. The head-shoulder width ratio is obtained as $r_{hs} = w_h / w_s$. The feature vector from an individual shape is defined as $\mathbf{v} = [r_a, r_{hs}, \phi_1, \phi_2, \phi_3]^T$.

It has been found by the inventors that the shape features are similar for human samples observed from camera tilt angles of the “small” and “near-zero” categories due to the nearly right angle of camera optical axis to the axis of human bodies. The measurements from the “large” category, however, are significantly different. Hence, human shapes are categorized into two classes corresponding to *large* and *small* (including near-zero) camera tilt angles divided at about 25° in the example embodiment. However, it will be appreciated that more classes can be used in different embodiments, including one class for each of the relevant camera tilt angle categories or ranges. A multivariate Gaussian is employed as the model for each class. The multivariate Gaussian models (i.e. $N_i(\bar{\mathbf{v}}_i, \Sigma_i), i=1,2$) can be obtained separately from offline training samples, where $\bar{\mathbf{v}}_i$ is the mean vector and Σ_i is the covariance matrix. Once the models have been obtained, they are used online (i.e. live) to estimate the camera tilt angle from observed human samples. For a human silhouette R represented by feature vector $\mathbf{v}$, the likelihood of R being observed from a camera tilt angle of the i^{th} class is

$$P_i(\mathbf{v}) \propto \frac{1}{|\Sigma_i|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{v} - \bar{\mathbf{v}}_i)^T \Sigma_i^{-1} (\mathbf{v} - \bar{\mathbf{v}}_i) \right] \quad (10)$$

Figure 5 shows shapes of human objects obtained from example images of various scenes. The top row shows images captured by the camera while the bottom row shows silhouettes of human objects obtained after a background subtraction of the respective scenes. The camera tilt angles progressively decrease from the left-most image to the right-most image. As can be seen from Figure 5, both the top and bottom positions of the silhouettes are widely distributed at large tilt angles, e.g. in image 502, but the top positions of the silhouettes become increasingly level as the tilt angle decreases, e.g. in image 504. A comparison between vertical extents of head and foot positions in the image recorded from a large number of observed humans may thus provide a global-level clue about the camera tilt angle.

In the example embodiment, one-dimensional (1D) histograms of vertical head and foot positions are generated. The data may be refined by truncating e.g. 10% tails of the histograms on both sides. The vertical ranges of the head and foot positions can be obtained as d_h and d_f , respectively and ratio $r_{hf} = d_h / d_f$ is calculated. Further, a linear fuzzy membership is employed to categorise the camera tilt angle from the value of r_{hf} obtained. Figure 6 shows membership functions for the 3 tilt angle categories with respect to the ratio of vertical distribution ranges of head and foot in which Q_j , where $j = 1, 2, 3$ are for large, small and near-zero tilt angle categories respectively. The parameters for r_{hf} ratios shown in Figure 6 may be determined empirically.

In the example embodiment, visual clues from shapes of objects and head/foot positions of a large number of samples are combined. By using the human shape model, samples of isolated and walking humans can be obtained. Let $R_t^j(x)$ be the j th 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 classes described above, i.e. $\exists i, P_i(\mathbf{v}_j) > T_p$, where T_p is a threshold probability, it is selected as one human sample. A large threshold T_p can be used to get good samples since abundant samples may be available.

After a sufficient number M (e.g. at least 100) of human samples have been collected, the likelihood for the j th category of camera tilt angles from visual clue of human shapes is calculated as $L_j = (1/M) \sum_m P_j(\mathbf{v}_m)$, where $j = 1, 2, 3$ and $P_3() = P_2()$. The likelihood for the j th category of camera tilt angles from visual clue of head/foot positions of the M samples can be obtained as Q_j . The category of the camera tilt angle is estimated as

$$a = \arg \max_{j=1,2,3} \{L_j + Q_j\} \quad (11)$$

With the estimated category of the camera tilt angle, the sample set of M observed humans is further filtered by the corresponding human shape model

$N_a(\bar{\mathbf{v}}_a, \Sigma_a)$. The selected N samples (where $N < M$) satisfying $P_a(\mathbf{v}_i) > T_p$ are used to determine the human perspective context of the scene.

Figure 7 shows a detailed flowchart 700 of a method to categorise a camera tilt angle for determining a perspective projection model. Such a method may be carried out online (i.e. live) when the surveillance system is started up. At step 702, background subtraction is applied to an image from the scene. At step 704, human samples are detected from the foreground regions (i.e. silhouettes). At step 706, if the silhouette is a human object, it is added to a sample set. At step 708, after a sufficient number of human samples are obtained in the sample set, the camera tilt angle is categorised into one of large, small or near-zero categories. At step 710, further sample filtering is carried out. At step 712, the perspective projection model is estimated from the human samples. At step 714, the perspective projection model is determined.

Data correction is applied into the process of model estimation for determining the human perspective context of the scene (step 712). The approach is an iterative process which contains two general steps in each iteration: model estimation (ME) and data tuning (DT). The first step generates a linear model of the perspective for humans from the updated sample data by using the least-square estimation, while the second step predicts and updates the sample data according to the obtained model. When the iterative process converges, the HPC model is obtained from predicted ideal sample data.

Let (μ, ν) be a pair of observed data from a person. For foot $\Leftrightarrow$ head homologies, the pair of data is the observed foot and head positions of a person in the image, i.e. $(\mu, \nu) = (\mathbf{x}_f, \mathbf{x}_h)$, and for scale $\Leftrightarrow$ position mappings, the pair of data is the observed vertical foot position and the height of the person in the image, i.e. $(\mu, \nu) = (y_f, h)$. Further, let the pair of the mappings between the observed data be $\mathbf{H}_1$ and $\mathbf{H}_2$, i.e.

$\nu = \mathbf{H}_1 \mu$ and $\mu = \mathbf{H}_2 \nu$. For foot $\Leftrightarrow$ head homologies, $\mathbf{H}_1 = H_{hf}$ and $\mathbf{H}_2 = H_{fh}$, and for scale $\Leftrightarrow$ position mappings, $\mathbf{H}_1 = \mathbf{m}_{sf}$ and $\mathbf{H}_2 = \mathbf{m}_{fs}$.

For initialisation, let the dataset be $\{\mu_i^j, \nu_i^j\}_{i=1}^N$ with $j = 0$ (i.e. the first iteration),
 5 and a weight ω_0 to control the data tuning.

Iteratively, estimates $\hat{\mathbf{H}}_1^j$ and $\hat{\mathbf{H}}_2^j$ of the linear transformation matrices are generated from the dataset $\{\mu_i^j, \nu_i^j\}_{i=1}^N$ using the least square estimation method. From the models $\hat{\mathbf{H}}_1^j$ and $\hat{\mathbf{H}}_2^j$ of the transformation, the predicted data are calculated as
 10 $\hat{\nu}_i^j = \hat{\mathbf{H}}_1^j \mu_i^j$ and $\hat{\mu}_i^j = \hat{\mathbf{H}}_2^j \nu_i^j$. Combining the observed and predicted data, the ideal data can be estimated as a weighted average such that

$$\begin{aligned} \mu_i^{j+1} &= (1 - \omega_j) \hat{\mu}_i^j + \omega_j \mu_i^0 \\ \nu_i^{j+1} &= (1 - \omega_j) \hat{\nu}_i^j + \omega_j \nu_i^0 \end{aligned} \quad (12)$$

In Equation (12), 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
 20 not become stable. The difference between the estimated data of the two latest iterations can be defined as

$$D_j = \frac{1}{N} \sum_i \left[(\hat{\mu}_i^j - \hat{\mu}_i^{j-1})^2 + (\hat{\nu}_i^j - \hat{\nu}_i^{j-1})^2 \right] \quad (13)$$

25 The iteration is terminated once a predetermined condition, e.g. $D_{j+1} \geq D_j$ (i.e. the difference has reached a minimum as a result of convergence), or $j > 5$, is satisfied. If the difference does not reach a minimum, the weight is adjusted as $\omega_j = \gamma \omega_{j+1}$ where $\gamma < 1$ (e.g. $\gamma = 0.7$), and j is set as $j+1$ for the next iteration. When the algorithm terminates, the refined samples are close to the virtual ideal positions. The linear transformation
 30 matrices obtained from the virtual ideal data become the perspective models for humans in the scene.

Figure 8A-8B show the performance of the ME-DT algorithm on synthetic data with various noise levels according to an example embodiment. In Figure 8A, the normalised error of the estimation (vertical axis) is plotted against the noise level (horizontal axis). Lines 802 (with points as '+' sign) and 804 (with points in circles) represent errors from the conventional approaches of standard least-square fitting, i.e. equivalent to the first step of ME-DT algorithm, and by removing 10% outliers according to the first estimation respectively. Both lines 802 and 804 are rapidly increasing at higher noise levels. In comparison, line 806 representing the errors from the ME-DT algorithm is relatively level even at very high noise levels, i.e. the method is robust against inaccurate data.

Figure 8B shows a series of plots corresponding the results of 5 iterative steps using the ME-DT algorithm of the example embodiment. The scattered points denote noisy synthetic training data. Lines 812 and 814 represent estimated mappings while line 816 represents the real model. As seen from Figure 8B, the scattered points quickly form a linear pattern as the iteration goes on, and lines 812 and 814 become increasingly close to the ideal line 816, i.e. the ME-DT algorithm can converge to the real model quickly even when it is applied on data of very high noise levels.

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 $\leftrightarrow$ 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 - head homology. Since H_{hf} is a mapping between homogeneous coordinates, it can be assumed that $a_{33} = 1$.

The left equation in (6) can be rewritten as

$$\begin{bmatrix} x_f & y_f & 1 & 0 & 0 & 0 & -x_f x_h & -y_f y_h \\ 0 & 0 & 0 & x_f & y_f & 1 & -x_f y_h & -y_f x_h \end{bmatrix} \mathbf{a} = \begin{bmatrix} x_h \\ y_h \end{bmatrix} \quad (14)$$

where $\mathbf{a} = [a_{11} \ a_{12} \ a_{13} \ a_{21} \ a_{22} \ a_{23} \ a_{31} \ a_{32}]^T$ are the rest of the elements of H_{hf} . Using Equation (14), the least square estimation $\hat{H}_{hf}$ can be obtained from a set of N samples $\{\mathbf{x}_{ht}, \mathbf{x}_{ft}\}_{i=1}^N$. Similarly, $\hat{H}_{fh}$ can also be obtained. The subsequent calculations follow as described above.

5

For a scene observed from a small camera tilt angle, the foot $\Leftrightarrow$ head homologies are still used to describe the human perspective context, but there are two points of difference in the implementation of ME-DT algorithm to estimate the homologies compared to that for *large* camera tilt angle. Firstly, 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 x-positions of the head and foot to the center point are caused by the variations of poses and segmentation errors. They often cause failure in homology estimation. Only the foot – head homology estimate $\hat{H}_{hf}^j$ is generated from the sample data using the least-square estimation.

15

As mentioned earlier, in y-direction, the head - foot homology is a linear mapping from a narrow range to a wide range. It may be difficult to obtain a stable mapping from noisy and inaccurate sample data. The head – foot homology is estimated by calculating the inverse of the foot – head homology, i.e. $\hat{H}_{fh}^j = (\hat{H}_{hf}^j)^{-1}$.

20

In the case of *zero* camera tilt angle, the scale $\Leftrightarrow$ position mappings in Equation (7) are used to describe the human perspective context. The sample data is a pair of foot position and the height of the body, i.e. $(\mu_i, \nu_i) = (y_{fi}, h_i)$. Through steps similar to those described above, least-square estimate is used to estimate $\hat{\mathbf{m}}_{fs}^j$ and $\hat{\mathbf{m}}_{sf}^j$.

25

Figure 9 is a flowchart 900 of an implementation of a perspective-constrained human detection method according to an example embodiment. At step 902, the camera tilt angle is categorised and the human perspective context is determined as described above, based on training samples of observed human objects. At step 904, once the

30

camera tilt angle is categorised, a corresponding human appearance model is selected for human detection. At step 906, after the HPC is determined, a scale-adaptive grid for human objects on the ground plane is generated. At step 908, a perspective-constrained human detection is carried out on a new input image obtained from the camera 910.

5

In the example embodiment, a Histogram of Oriented Gradient (HOG) human detector is employed. To adapt to great scale variations of humans in some scenes, the detection window is divided into 4×8 cells where the size of the cells varies from 4×4 pixels to 20×20 pixels resulting in detection windows of 17 scale levels from 16×32 to 80×160 pixels. In addition, each detection window consists of 3×7 blocks formed by 2×2 cells in a sliding fashion.

10

Using learned HPC, a scene-adaptive grid on the ground plane which determines the scale and orientation of a human in a position in the image can be generated. Figure 10 shows a series of images of various scenes (upper row) and the corresponding grids generated thereon (lower row), having foot grid points 1002 and head grid points 1004. The grid is generated row by row from the bottom of the image which is the closest position to the camera on the ground plane. The distance between adjacent grid points is preferably about half of an average human width. The vertical distance between two adjacent rows is set such that there is $\beta\%$ vertical overlap of human heights in the two rows, e.g. $\beta = 85\%$.

20

The generation of grid points may be terminated when the human height calculated for the next row is less than the minimum height of human objects which can be detected in the image, or when the position of head at a row reaches the upper boundary of the image. In some embodiments, the vertical vanishing point may be determined from the samples. As the grid density is determined by the camera tilt angle and sizes of human object in the scene, the number of grid points, hence the number of detection windows, is not much related to the image resolution.

25

30

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 human object and denoted as $\mathbf{x}_f$. The corresponding head position $\mathbf{x}_h$ can then be estimated using the learned HPC. The height for the possible human object can be calculated as $h = \|\mathbf{x}_f - \mathbf{x}_h\|$, which is
 5 quantified into one of the 17 levels as described above. Hence, the size of the detection window can be determined. The orientation of the window is

$$\theta = \tan^{-1} \left(\frac{x_h - x_f}{y_h - y_f} \right)$$

with respect to the vertical axis of the image. The sub-region within the detection window
 15 is rotated to the upright position and the HOG feature vector from the window can be obtained. The feature vector is then fed to a support vector machine (SVM) for classification. The output is the probability defined in Equation (5). If the output value is greater than a threshold T_h , a person is detected at the grid point.

20 The human detection method according to the example embodiment may significantly reduce the number of detection windows for each scene. Conventionally, with no knowledge of perspective constraints on human appearance in the scene, it is necessary to scan the image with the detection windows of multiple scales. Normally, for an image of 320 x 240 pixels, over 200,000 windows of 5 scales are used to adapt to the
 25 scale variations of human objects in a natural scene. Using the method of the example embodiment, with the perspective constraints about human objects in a scene, there is no need to detect human objects at positions which may be too high over the ground plane. In addition, even in scenes where the grids cover almost all the image, the use of detection windows of multiple scales for each position can be avoided.

30 The number of detection windows and the reduction rates for a number of different scenes using the method of the example embodiment are shown in Table 2. On average, the detection windows generated according to the human perspective context is less than 1% of those used by dense scanning for small resolution images and is

about 0.1% for large resolution images. This means the method of the example embodiment can detect humans up to 100 times faster than conventional methods.

Table 2

Scene No.	Image Size	No. of windows	Reduction Rate	Tilt Angle
1	320 x 240	363	0.053%	zero
2	360 x 288	511	0.074%	zero
3	320 x 256	244	0.036%	zero
4	176 x 144	938	1.32%	small
5	176 x 144	828	1.17%	small
6	176 x 144	489	0.69%	small
7	360 x 288	1564	0.228%	large
8	384 x 288	1826	0.266%	large
9	184 x 144	513	0.723%	large
10	176 x 144	732	1.03%	large
11	320 x 240	1285	0.187%	large
12	192 x 144	707	0.997%	large
13	192 x 144	774	1.092%	large
14	176 x 144	356	0.502%	large

5 Furthermore, another advantage of applying the perspective context to human detection in the example embodiments is the reduction of the chances to falsely detect human-like objects in the scenes. The dominant features of standing and walking persons include the boundary lines on both sides. In natural scenes, many objects, such as pillars, tree trunks, windows, and doors, can look like standing human objects and thus result in a false positive in conventional detection systems. Using the human perspective context of the scene in the example embodiments, such objects are, however, often either too high or too short with respect to human metric. Thus, at the corresponding positions in the image, false positive detection as human objects can be avoided. Some examples of such situations are shown in Figure 11. The first row shows the human detection results by using a conventional dense scanning methodology. It is

noted, that in order to suppress too many false positives, already a very high threshold was used to the outputs of a SVM detector, *i.e.* $T_h = 1.1$

The second row shows the results of perspective constrained human detection according to an example embodiment, and with a small threshold $T_h = 0.3$. As can be seen from Figure 11, the results from the conventional method still contain a number of false detections 1102 which may be e.g. trees or pillars. On the other hand, the results from the method of the example embodiment correctly show all human objects in the scene without false detections. Thus, it can be observed that several false positives detected by using a conventional dense scanning methodology will not appear when the human perspective contexts of the scenes are applied in example embodiments, even though a lower threshold is used.

On the other hand, evaluations on 15 example comparative sequences also showed that, with a deliberately similar false positive alarm rate per-frame (e.g. 2 FPPF) as obtainable for conventional methods, the detection rate of human objects using the example embodiments can be improved from about 80% to about 91.2%.

It will be appreciated by a person skilled in the art that numerous variations and/or modifications may be made to the present invention as shown in the specific embodiments without departing from the spirit or scope of the invention as broadly described. The present embodiments are, therefore, to be considered in all respects to be illustrative and not restrictive.

CLAIMS

1. A method of learning human perspective context (HPC) for human detection in video data, the method comprising the steps of:

5 selecting a plurality of samples based on human detection from the video data;

categorizing a camera tilt angle into one of two or more categories based on the selected samples; and

learning the HPC from the selected samples;

10 wherein the samples are filtered based on the categorization of the camera tilt angle prior to learning the HPC, a model calculation for the HPC is based on the categorization of the camera tilt angle, or both.

2. The method as claimed in claim 1, wherein the selecting of the
15 samples comprises

detecting foreground objects in the video data;

determining probabilities of respective ones of the foreground objects corresponding to one of at least two shape feature models; and

20 selecting respective objects as one of the samples if either probability meets a first threshold.

3. The method as claimed in claim 2, wherein the determining of the probabilities of the respective ones of the foreground objects belonging to one of the shape feature models is based on a shape feature vector of the foreground object.
25

4. The method as claimed in claims 2 or 3, wherein the categorizing of the camera tilt angle comprises determining head and foot positions of the respective selected samples, and determining likelihoods for the plurality of selected samples belonging into respective ones of two or more classes of head and foot
30 positions corresponding to the respective categories of the tilt angle.

5. The method as claimed in claim 4, wherein the categorizing of the camera tilt angle further comprises determining said one category based on both the likelihoods for the plurality of selected samples belonging into respective ones of the

two or more classes of head and foot positions and the probabilities of respective ones of the foreground objects corresponding to one of at the least two shape feature models.

5 6. The method as claimed in any one of claims 2 to 5, wherein the filtering of the samples comprises filtering by one of the shape feature models corresponding to the categorization of the camera tilt angle.

10 7. The method as claimed in any one of claims 2 to 6, wherein the model calculation for the HPC uses either foot $\Leftrightarrow$ head homologies or scale $\Leftrightarrow$ position mapping to describe the HPC depending on the categorisation of the camera tilt angle.

15 8. The method as claimed in claim 7, wherein the model calculation further uses different implementations of describing the HPC using the foot $\Leftrightarrow$ head homologies depending on the categorisation of the camera tilt angle.

20 9. The method as claimed in any one of the preceding claims, further comprising applying the learnt HPC in human detection from subsequent video data.

25 10. The method as claimed in claim 9, wherein the applying of the learnt HPC comprises generating a scene-adaptive grid which defines a scale and orientation of a human in a position in an image, and limiting detection windows of a human detection algorithm based on the scene-adaptive grid.

 11. The method as claimed in claim 10, wherein the human detection algorithm is selected from two or more algorithms based on the categorization of the camera tilt angle.

30 12. The method as claimed in claims 7 or 8, wherein the model calculation for the HPC, either foot $\Leftrightarrow$ head homologies or scale $\Leftrightarrow$ position mapping, comprises using a model estimation - data tuning (ME-DT) algorithm.

13. A system for learning human perspective context (HPC) for human detection in video data, the system comprising:

means for selecting a plurality of samples based on human detection from the video data;

5 means for categorizing a camera tilt angle into one of two or more categories based on the selected samples; and

means learning the HPC from the selected samples;

wherein the means for learning the HPC filtered the sample based on the categorization of the camera tilt angle prior to learning the HPC, uses a model
10 calculation for the HPC based on the categorization of the camera tilt angle, or both.

14. The system as claimed in claim 13, wherein the means for selecting of the samples

detects foreground objects in the video data;

15 determines probabilities of respective ones of the foreground objects corresponding to one of at least two shape feature models; and

selects respective objects as one of the samples if either probability meets a first threshold.

20 15. The system as claimed in claim 14, wherein the determining of the probabilities of the respective ones of the foreground objects belonging to one of the shape feature models is based on a shape feature vector of the foreground object.

25 16. The system as claimed in claims 14 or 15, wherein the means for categorizing the camera tilt angle determines head and foot positions of the respective selected samples, and determines likelihoods for the plurality of selected samples belonging into respective ones of two or more classes of head and foot positions corresponding to the respective categories of the tilt angle.

30 17. The system as claimed in claim 16, wherein the means for categorizing the camera tilt angle further determines said one category based on both the likelihood of the likelihoods for the plurality of selected samples belonging into respective ones of the two or more classes of head and foot positions and the

probabilities of respective ones of the foreground objects corresponding to one of at the least two shape feature models.

18. The system as claimed in any one of claims 14 to 17, wherein the means for learning the HPC filters the samples by one of the shape feature models corresponding to the categorization of the camera tilt angle.

19. The system as claimed in any one of claims 14 to 18, wherein means for learning HPC uses either foot $\Leftrightarrow$ head homologies or scale $\Leftrightarrow$ position mapping to describe the HPC depending on the categorisation of the camera tilt angle.

20. The system as claimed in claim 19, wherein the means for learning the HPC used different implementations of describing the HPC using the foot $\Leftrightarrow$ head homologies depending on the categorisation of the camera tilt angle.

21. The system as claimed in any one of claims 13 to 20, further comprising means for applying the learnt HPC in human detection from subsequent video data.

22. The system as claimed in claim 21, wherein the means for applying the learnt HPC generates a scene-adaptive grid which defines a scale and orientation of a human in a position in an image, and limits detection windows of a human detection algorithm based on the scene-adaptive grid.

23. The system as claimed in claim 22, wherein the human detection algorithm is selected from two or more algorithms based on the categorization of the camera tilt angle.

24. A data storage medium having computer code means for instructing a computing device to execute a method of learning human perspective context (HPC) for human detection in video data, the method comprising the steps of:

selecting a plurality of samples based on human detection from the video data;

categorizing a camera tilt angle into one of two or more categories based on the selected samples; and

learning the HPC from the selected samples;

- 5 wherein the samples are filtered based on the categorization of the camera tilt angle prior to learning the HPC, a model calculation for the HPC is based on the categorization of the camera tilt angle, or both.