Deformable Parts Model for People Detection in Heavy Machines Applications

Manh-Tuan Bui, Vincent Frémont, Djamal Boukerroui, Pierrick Letort

Deformable Parts Model for People Detection in Heavy Machines Applications

Manh-Tuan Bui, Vincent Frémont, Djamal Boukerroui
Université de Technologie de Compiègne (UTC)
CNRS Heudiasyc UMR 7253
Compiègne, France
Email: manh-tuan.bui@utc.fr

Pierrick Letort
Technical center for the Mechanical Industry (CETIM)
Senlis, France

***Abstract*—In this paper we focus on the evaluation of the deformable part model (DPM) proposed by Felzenszwalb et al. [10] in the context of vision-based people detection in heavy machines applications. The proposed system uses a single fisheye camera to provide a wide field-of-view (FOV) at low cost. However, the fisheye optical distortions present several difficulties for image processing and object recognition. The DPM approach shows important flexibility when dealing with varying object's form. It gives good performances on people detection when images present strong fisheye distortions. Base on the analysis of DPM in the context of fisheye image, we proposed an adaptive detector which is more suitable.**

***Index Terms*—Heavy machines, pedestrian detection, deformable part model, fisheye images, histogram of oriented gradients, latent support vector machine.**

I. Introduction

Construction sites are considered as a high risk working environment. People who work near heavy machines are constantly at risk of being struck by a working machine or its components. Accidents between machines and people represent a significant contribution to construction health and safety hazards. It is hard for drivers to keep watching all around their vehicle and fulfill their productive task at the same time, due to the complicated shapes of these machines. It is therefore mandatory to develop an advanced driver assistance systems (ADAS) to help the driver monitor the surrounding area and being able to raise a pertinent alarm when people are threatened. Therefore the main required functionality by such ADAS is the people detection. Notwithstanding many years of progress, safety system for people working around heavy machine is still an unresolved issue.

To solve the problem of safety on heavy machine, various kinds of sensors have been tested and compared, individually or combined. Range sensors, like radar, Light Detection And Ranging (Lidar) and ultrasonic [1], which have good performance in detecting obstacles, are usually unable to distinguish between objects and people. Heavy machines often work in complicated terrains with a lot of nearby objects. In these situations, range sensors will trigger a permanent alarm, which is useless and annoying for the drivers. The last commonly used sensor is the camera. It offers the best option as a low-cost and polyvalent sensor. Moreover, computer vision and image processing tools provide the ability to recognize various kinds of objects, including people.

Recently, people detection systems on automobile, which share a lot of characteristics with the context of heavy machines, has known important progresses. Although the problematic is similar in both contexts, we can clearly distinguish between the two. In the automobile field, cars need to stop if there is an obstacle, no matter if it is a pedestrian or an object. The task of recognizing people is more important for heavy machines where the main requirement is human's safety. Besides, cars often operate at a higher speed and on straight ways. While it is important for the system on automobile to be able to detect people at far distances, heavy machines need a larger field of view (FOV) to cover the nearby area. Construction machines often have a complicated shape and large size, which can also benefit from the large FOV. This reason encourages us to use fisheye camera as the main sensor in our system. Objects detection in fisheye images is challenging. Unlike in perspective images, the appearance of objects captured by these cameras is strongly distorted. Wrapping the image into a local perspective image is the direct way to avoid non-perspective deformations. Unfortunately, besides adding computational load, this approach also creates undesirable effects. Daniilidis et al. [6] and Bülow [4] are the first researchers who argued that the warping of wide-angle images should be avoided. Recently, there are others researchers who proposed approaches to increase matching rate of Scale-invariant feature transform (SIFT) for wide-angle images [11], [15].

On the other side, the Deformable Parts Model (DPM) of [10] and its variants have recently gained a lot of attentions in object detection and recognition. The approach performs especially well in detecting people in hard conditions. Indeed, they are the winners of some recent Pascal-VOC detection challenges [8], [9] and have obtained the best scores in others benchmark [7]. This approach can be considered as the current state of the art methods. The DPM can represent an object model with different parts floating around their reference locations and finding the optimal part-configuration at every root position. This elegant way of representation brings a lot of benefit in detecting a person in different postures.

Our contribution in this paper lies in the evaluation of the DPM approach in the context of strong radial distortions in fisheye images. The analysis focus on the influence of different

types of training dataset and in comparing the images features used in the DPM approach with the histogram of oriented gradient of Dalal and Triggs [5]. A fisheye images dataset in the context of heavy machine has also been created for testing purposes. A DPM-based people detector, adapted to fisheye images is proposed as the result of all the analysis. The paper is organized as follows. First, a brief description of the DPM is presented in section II. Section III discusses the model of fisheye camera and proposes an adaptive DPM-based detector. The evaluation results are presented in section IV. Conclusions and perspectives of our work will be presented in section V.

II. The Deformable Part Model

We are mainly interested in 2 contributions proposed in the DPM framework: the feature vector reduction method and the deformable model trained by the latent SVM method. This section will recall the main ideas and clarify some details of these two methods in our research.

A. Histogram of oriented gradient dimensionality reduction

Histogram of oriented gradient (HOG) is undoubtedly the most used image feature in people recognition. The feature used in the DPM is a derived version. In this approach, the authors use the Principal Component Analysis (PCA) to justify the reduction of dimension of the HOG vector. In this paper, we call it *reduced-HOG* feature to distinguished from the classical HOG of Dalal and Triggs [5], denoted as *conventional-HOG*. Most of the parameters of the *reduced-HOG* are the same as the *conventional-HOG*. One cell includes 8×8 pixels, one block includes 4 cells and the number of angular bins of the vector gradient is 9.

Let $C(i,j)$ the feature vector corresponding to the histogram of one cell before the block normalization. One cell belongs to 4 blocks around it (Fig. 1 left). For each block, the HOG vector of 9 cells $C(i,j)$ is normalized with a different factor called gradient energy of the block $N_{\delta,\gamma} = (\|C(i,j)\|^2 + \|C(i+\delta,j)\|^2 + \|C(i,j+\gamma)\|^2 + \|C(i+\delta,j+\gamma)\|^2$ with $\delta,\gamma \in \{-1,1\}$. The *conventional-HOG* vector represents one cell after normalization and is thus a 36-dimensional vector, where each 9 dimensions corresponds to one normalization block (Fig. 1 and Eq. 1).

$$H(i,j) = \begin{pmatrix} H(i,j)_{-1,-1} \\ H(i,j)_{+1,-1} \\ H(i,j)_{+1,+1} \\ H(i,j)_{-1,+1} \end{pmatrix} = \begin{pmatrix} C(i,j)/N_{-1,-1}(i,j) \\ C(i,j)/N_{+1,-1}(i,j) \\ C(i,j)/N_{+1,+1}(i,j) \\ C(i,j)/N_{-1,+1}(i,j) \end{pmatrix} \quad (1)$$

By applying PCA, Felzenszwalb et al. [10] proposed that the 36-dimensional feature vector of one cell can be reduced to 11 without sacrificing important information. For the reason of simplicity, a method of projection to reduce the dimension of HOG vectors has been proposed (Fig. 1-right). The reduced vector includes 9 dimensions that correspond to 9 angular-bins and 4 dimensions that reflect the overall gradient energy in different areas around the cell $C(i,j)$. The 9 bins in the HOG vector are normally contrast-insensitive. Felzenszwalb et al. suggest empirically that the detection performance on some

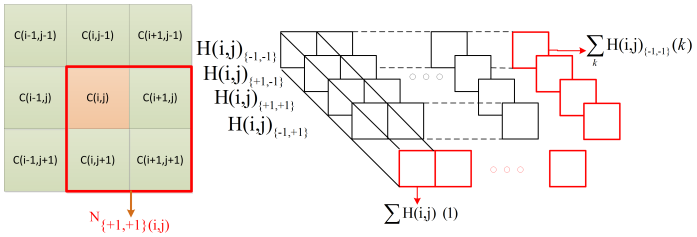

Figure 1: The dimensional-reduced-HOG computation process: Normalization blocks around one cell in HOG (left). Transformation projection to reduce dimension of the dimensional reduced HOG (right).

object categories is improved using contrast-sensitive features, while other benefits from contrast-insensitive information [5]. Therefore they suggested the use of 31 dimensions feature vector (9 bins contrast-insensitive + 18 bins contrast-insensitive + 4 gradient energy).

B. Deformable part model

The deformable parts model is the most important contribution of the DPM over the conventional HOG. In the simple model of the conventional-HOG, the score of the detector can be considered as a scalar product between the feature map $\Phi(x)$ and a vector of parameter β:

$$f_\beta(x) = \boldsymbol{\beta}.\Phi(x) \quad (2)$$

where β is defined through the training process.
In the part-based model:

$$f_\beta(x) = \max_{x \in Z(x)} \boldsymbol{\beta}.\Phi(x,z) \quad (3)$$

where the latent variable $\boldsymbol{z} = (p_0, p_1, ..., p_n)$ with $\boldsymbol{p_i} = (x_i, y_i, l_i)$ is the specification of i^{th}-part configuration at the position x_i, y_i and scale l_i. The purpose of the training process in the latent SVM process is also to learn the vector of the model parameters $\boldsymbol{\beta} = (\boldsymbol{F_0}, \boldsymbol{P_1}, ..., \boldsymbol{P_n}, b)$ where $\boldsymbol{F}_0$ represents the root filter model and $\boldsymbol{P_i} = (\boldsymbol{F_i}, \boldsymbol{v_i}, \boldsymbol{d_i})$ corresponds to the parameters of different parts. $\boldsymbol{F_i}$ is denoted as the filter model of part i, $\boldsymbol{v_i}$ is the anchor position of i^{th}-part in the image frame and $\boldsymbol{d}_i$ is the deformation features, a 4D-vector of coefficients, corresponding to x, y, x^2, y^2.

The idea behind the deformable part model is to represent the object of interest model using a lower-resolution "root" template $\boldsymbol{F}_0$, and a set of spatially flexible higher-resolution "part" template $\boldsymbol{P_i} = (\boldsymbol{F}_i, \boldsymbol{v}_i, \boldsymbol{d}_i)$. Each part captures local appearance properties of an object, and the deformations are characterized by the deformation costs. The deformation cost is calculated by a quadratic equation characterized by 4 coefficients $\boldsymbol{d}_i$. This equation and the optimal positions of each part in the model are obtained through the training process. Once the vector $\boldsymbol{\beta}$ is defined, the score of one object hypothesis can be computed as [10]:

$$score(p_0, p_1, ..., p_n) = \sum_{i=0}^{n} \boldsymbol{F_i}.\phi(H, p_i) - \sum_{i=1}^{n} \boldsymbol{d}_i.\phi_d(dx_i, dy_i) + b \quad (4)$$

where H is the feature pyramid and $\phi(H, p_i)$ denotes the vector obtained by concatenating the feature vectors in the format of i^{th}-part filter. $(dx_i, dy_i) = (x_i, y_i) - (\lambda(x_0, y_0) + \boldsymbol{v}_i)$

gives the displacement if the i^{th}-part relative to its anchor position and the deformation features are typically $\phi_d(dx, dy) = (dx, dy, dx^2, dy^2)$.

III. DPM ON FISHEYE IMAGE

A. Calibration model and field of deformation calculation

Wide-angle cameras have noticeable geometric distortions. While these distortions may be artistically interesting, it is generally desirable to remove them in many applications in computer vision. The geometric distortions include two major components: radial and tangential. Radial distortion causes image points to be translated by an amount proportional to their radial distance to the optical center. Tangential distortions (or decentering distortion) are generally less significant than radial distortions and are produced by the misalignment of the optical centers of various lens elements.

Given a real point $\boldsymbol{P} = (X, Y, Z)^T$, the undistorted point projected on the image sensor will be represented as $\boldsymbol{p} = (u, v)^T = \left(X\,\frac{f_x}{Z}, Y\,\frac{f_y}{Z}\right)^T$ with f_x and f_y the focal lengths of the camera optic. In the case of a wide-angle camera, the position of the distorted point on the image is given by:

$$\tilde{\boldsymbol{p}} = \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} = \boldsymbol{p} + \delta\boldsymbol{p} + \boldsymbol{p_0} \tag{5}$$

where $\delta\boldsymbol{p} = \begin{pmatrix} \delta u \\ \delta v \end{pmatrix} = \begin{pmatrix} \delta u^{(r)} + \delta u^{(t)} \\ \delta v^{(r)} + \delta v^{(t)} \end{pmatrix}$ is the approximated distortion and $\boldsymbol{p_0} = (u_0, v_0)$ is the principal point of the camera. $\delta u^{(r)}$, $\delta v^{(r)}$ represent radial distortions and $\delta u^{(t)}$, $\delta v^{(t)}$ are the tangential distortions along the two image axes.

Among different distortion models, the standard polynomial model is the most popular [13]. In this paper, the standard polynomial model of third degree is used:

$$\begin{pmatrix} \delta u^{(r)} \\ \delta v^{(r)} \end{pmatrix} = \begin{pmatrix} \tilde{u}(k_1 r_d^2 + k_2 r_d^4 + k_3 r_d^6) \\ \tilde{v}(k_1 r_d^2 + k_2 r_d^4 + k_3 r_d^6) \end{pmatrix} \tag{6}$$

$$\begin{pmatrix} \delta u^{(t)} \\ \delta v^{(t)} \end{pmatrix} = \begin{pmatrix} 2p_1\tilde{u}\tilde{v} + p_2(r_d^2 + 2\tilde{u}^2) \\ p_1(r_d^2 + 2\tilde{v}^2) + 2p_2\tilde{u}\tilde{v} \end{pmatrix} \tag{7}$$

The optimal values of (f_x, f_y, u_0, v_0) and $(k_1, k_2, k_3, p_1, p_2)$ are estimated through a calibration process. It is a prerequisite for any accurate geometric measurements from image data. Most of calibration methods use known geometric patterns, such as corners, dots, circles, lines and other features that can be easily extracted from images. The method described in [12] and implemented in [2] was used in our work.

B. Adaptive DPM-based model

Two main remarks can be drawn from the observation of the appearance of a person in fisheye images:

- The distortion in wide-angle camera is not identical over all the image area. It is particularly strong at close range and at the image boundaries.
- The distortion of body's part is minor compared to the full-body appearance.

Starting from these observations and our understanding of the DPM (described in section II-B), we propose the following improvements:

- Since the DPM approach has good performance on people detection on perspective images, it should have equally good performance in detecting the person at a long distance to the camera in the fisheye context.
- In the extreme-cases, when people are staying very close and at the border of the camera FOV, we will have low response of the root-filters F_0 but equally high response on part-filter F_i. We could adjust the deformation features $\boldsymbol{d_i}$ and $\boldsymbol{v}_i$ (see Eq. 4) to adapt the deformable part model to the radial distortion. This adaptation will depend on the position of the object in the FOV of the camera. We have also noticed that the radial distortions have minor effects on small local region because the relative dislocation between neighbor-points is small. Once the anchor positions $\boldsymbol{v_i}$ is well defined, the deformation features d_i are not sensitive to distortions.

The second assumption has been developed into an adaptive DPM-based approach where we relocate the anchor position of each part in the deformation model. Wide-angle optics are very different from one to another. Given the height h_c of the camera, an approximate size $(W \times H)$ of a person and his position angle θ and distance d to the camera, the center point of the person is given by

$$\boldsymbol{P_0} = \begin{pmatrix} X_0 \\ Y_0 \\ Z_0 \end{pmatrix} = \begin{pmatrix} d\,\sin\theta \\ H - h_c \\ d\,\cos\theta \end{pmatrix} \tag{8}$$

The center point $\boldsymbol{p_i}$ of the i^{th} part in the model on a perspective image is defined by $\boldsymbol{p}_i = \boldsymbol{p_0} + \boldsymbol{v_i}$. Given a trained DPM model with the root filter of size $(w \times h)$, the center position $\boldsymbol{p_i}$ of each part can be converted into a 3D position as

$$\boldsymbol{P}_i = \boldsymbol{P_0} + \boldsymbol{v_i}. \begin{pmatrix} W/w \\ H/h \\ 1 \end{pmatrix} \tag{9}$$

Both points $\boldsymbol{P_0}$ and $\boldsymbol{P}_i$ can be projected on the fisheye image using the camera projection equation and a given distortion model as $\tilde{\boldsymbol{p}}_0$ and $\tilde{\boldsymbol{p}}_i$ (section III-A). The distortion of the fisheye optics can therefor be taken into account by an estimate of the displacement of each part relatively to the root part. We define $\Delta\tilde{\boldsymbol{v}}_i = \tilde{\boldsymbol{p}}_i - \tilde{\boldsymbol{p}}_0$. The score of one object hypothesis (Eq. 4) in the adaptive model will have $(dx_i, dy_i) = (x_i, y_i) - (\lambda(x_0, y_0) + \boldsymbol{v_i} + \boldsymbol{\Delta\tilde{v}_i})$ gives the displacement if the $i^{th}-$ part relative to its anchor position. The displacement of each part $\Delta\tilde{\boldsymbol{v}}_i$ depends on the position of the person in the camera coordinate so it evolves in function of the position of root filter and the scale in the feature pyramid. Value of $\Delta\tilde{\boldsymbol{v}}_i$ can be computed offline and saved into lookup tables. In online detection, this approach intervenes in the final score computing using responses from the root by adding one lookup table operation. The computing cost of this adaptation is actually insignificant.

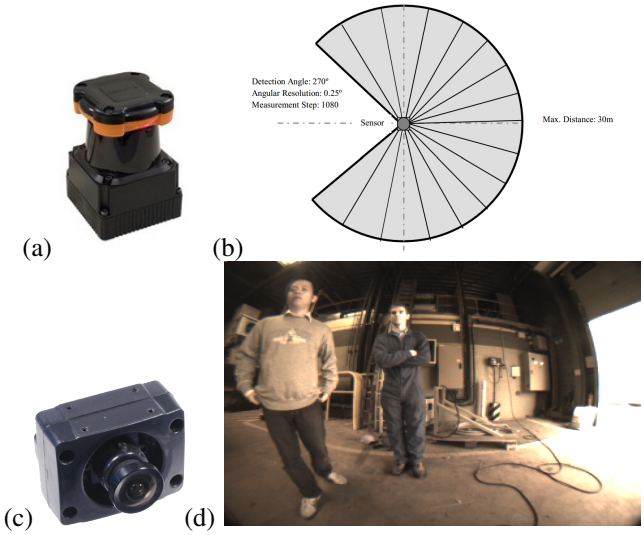

Figure 2: Our proposed sensors system. The Lidar Hokuyo UTM-30LX (a) and its horizontal field of view (b). The camera PointGrey Firefly MV (c) and its example image (d)

IV. Analysis, Experiments and Results

The dataset and the evaluation protocol are presented in sections IV-A and IV-B. Our analysis on testing the DPM approach in the context of fisheye images, begin with modifying the training database. The mix-training-dataset method [3] combined with the DPM are also taken into consideration. The dimensional-reduced-HOG is thorough evaluated in term of computational resource and performance. We also present the result of the adaptive DPM-based approach proposed in section III.

A. The dataset

Datasets take a very important role in the process of the development of a detection algorithm. A well defined dataset is not only useful for evaluating the approach but it also takes part in the training process to improve the performance of the detector. To the best of our knowledge, there are no others available dataset which provides at the same time synchronized fisheye images and LIDAR data. The LIDAR data provide precisely the distance of all objects present in the FOV. These data can be combine with the image sequence to lower the computation time of the detection algorithm and to reduce the false detection rate. The heavy machine context has some characteristics: outdoor light conditions, strong vibrations, and the presence of brutal shocks might make the detection process much harder. There are essential needs for a new dataset in order to identify conditions under which current detectors fail and focus the effort on these difficult cases.

The proposed dataset consists of images captured from one fisheye camera (*Point Grey Firefly MV USB2.0*), one conventional camera (*Sony PlayStation Eye for PS3*) and one range-sensor (*LIDAR Hokuyo UTM-30LX-EW*). The used sample frequencies are $10Hz$ for both cameras and 40Hz for the LIDAR (Fig. 2).

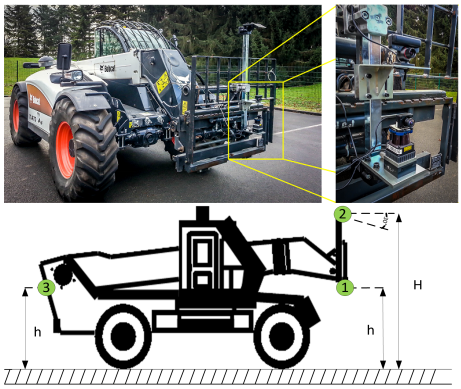

Figure 3: Top: Real image of the acquisition system setup in the configuration 1. Bottom: Sensor configurations map: $h = 110cm$ and $H = 210cm$

These data are taken on board of a telescopic-forklift, namely a Bobcat-TL470, as shown in Fig. 3 (a professional driver has been recruited to operate the machine during the experiments). In most of the data sequences, the machine moves on a clear terrain. The scenarios defined within the experiment aim to simulate frequently meet situations on a construction site. People wore different kind of clothes, including helmet, reflective vests and civil clothes. Different situations of occlusion were also simulated. We defined 3 configurations sensors' positions denoted by 1, 2 and 3 as illustrated in Fig. 3. In the first configuration, all the sensors are in front of the machine at position 1. In configuration 2, we move the fisheye camera to position 2, pointing down to the ground with an angle of 30°. In the last configuration, the three sensors are positioned at the rear of the machine, at position 3.

The experiments are divided into 4 setups and 6 scenarios. All of the scenarios are pre-defined and took place under security control. The first three setups correspond to the three configurations of the system (Fig. 3). The acquisitions in the last setup were performed after sunset in order to simulate very low light conditions. The main source of light comes from the headlights of the machine in this experiments. In these lighting conditions, the reflective safety clothes appears with a high contrast on the image. Therefore it is an important image cue.

B. Evaluation protocol

The detection system takes an image and returns bounding boxes with corresponding scores or confidence indicators. A detected bounding box A and a ground truth bounding box B form a match if they have a sufficient overlap area. In the PASCAL challenge [9], the overlap criterion between the two bounding box A and B is $t = \frac{A \cap B}{A \cup B} > t_0$ where t_0 is a threshold. $t_0 = 0.5$ is considered reasonable and is commonly used. The protocol of evaluation is adapted from the tool of Dollár which was used in [7]. As the context of heavy machines requires reducing the false detection rate, the results are presented in miss rate against false positive per image (FPPI). Only bounding boxes with a height more than 50 pixels are considered in the evaluation. Each detected bounding box

Training dataset	Positive samples	Negative sample
Inria	1371 samples from 614 images	1218 images
Inria-mix	1371 samples from 614 images	1218 images
fisheye	1520 samples from 810 images	1202 images

Table I: Training datasets characteristic

may be matched once with the ground truth and redundant detections are considered as false positives (FP).

We plot miss rate versus FPPI (lower curves indicate better performance) by varying the threshold of the detector. Decreasing the threshold level results in reducing the miss rate but in increasing the false positive per image rate. At a given FPPI, it is easy to compare the performance of different detectors. The log-average miss rate is used to summarize detectors' performance. The log-average miss rates is computed by averaging miss rate at nine FPPI evenly spaced in log in the range 10^{-2} to 10^{0} (for curves that end before reaching a given FPPI rate, the minimum miss rate achieved is used). When curves are somewhat linear in this range, the log-average miss rate is similar to the performance at 10^{-1} FPPI [7], [14]. The displayed legend entries are ordered by log-average miss rate from the worst to the best.

In the scope of our analysis, 3 different training datasets were taken into consideration: Inria [5], Inria-mix and Fisheye (see table I). The fisheye images (presented in section IV-A) are partially annotated and split into a train set and a testing one. The annotation for the ground truth of these image sequences are done by the labeling tool of Dollár et al. This tool requires the bounding box around objects in some key-frames and provides linear interpolation to infer the bounding boxes of the same object in intermediate frames. The objects can be labeled, in our case as: "person", "person sitting" and "occluded". The Inria-mix set have been build based on the Inria set. Starting from a training dataset without distortions, we replaced randomly 50% of undistorted images by distorted ones. The latter are simulated at different distortion angles. The total number of positive and negative sample images in the new training dataset is the same as the original one. For more details about the process of artificially generating distorted image samples, the reader is referred to [3].

The notation of different detectors used in this paper is based on the name of the dataset used for training, the name of the features extraction method and the name of the detection algorithm. The testing dataset has 7 image sequences of 5747 fisheye images.

C. Results

a) The conventional HOG vs the dimensional reduced HOG : In terms of computational cost, the reduction of the feature vector dimension improves slightly the speed of the DPM. From 36 to 32-dimentional vector, the gain is not significant. As a result, we observed a gain of about 8% in detection time (measured using 100 images at resolution VGA 480×640 on a PC).

The results shown in Fig. 4 are inconsistent. While the conventional-HOG feature works better with Inria and Inria-mix training dataset, the reduced-HOG features seems to work better when fisheye images are used for training. The combination of contrast-insensitive and contrast-sensitive features does not give any benefit in the general case. Contrast-sensitive features may be helpful when the object to background contrast does not change sign. This is the case for example, when the background color is always darker than the reflective-vest. The goal of the DPM approach is to detect numerous kind of objects so combining contrast-sensitive with contrast-insensitive in the feature vector is a relevant solution.

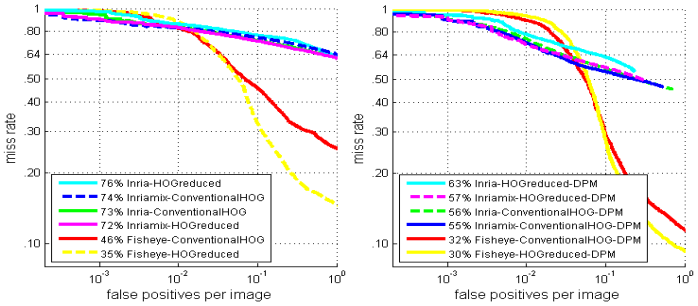

Figure 4: Detection performance of the dimensional-reduced-HOG vs the conventional-HOG using the root-filter only (left) and using the DPM approach (right).

b) The influences of the training dataset: The first experiment involves the DPM approach trained with 3 different datasets: Inria, Inria-mix and Fisheye Dataset. The differences between the model trained with Inria and Inria-mix is not visually noticeable (Fig. 6) but there is a small boost of performance on the detection results (Fig. 5 left). This result conforms with the conclusion in [3]. Enriching the training dataset can handle the distortion on the people's appearance, even with the DPM approach.

Unfortunately, there is a trade-off between the image quality and the amount of distortion added to the training examples. This process has the drawback of introducing missing pixels that have to be filled by interpolation. This phenomenon is proportional to the amount of distortion and it has bad effects on the performance of the detector. In practice, image samples simulated at an angle superior to 60° are unusable because the samples loose most of image details. The result of Inria-mix detector in this experiment is the best that we can obtain from the mix-training-dataset approach. The degradation of image quality during the distortion process affects remarkably the performance of the detection. The model trained by 100% distorted sample images gives an unrecognizable person model and in experiment, the detector does not work at all.

In the perspective camera model, the HOG responses are strong on silhouette contours, especially the head, shoulders and feet [5]. Based on this observation, the DPM defines the anchor position of parts at the high respond region of the root filter. In the case of the fisheye detector model, although we can recognize the silhouette contours of a person but there are very little correlation between the perspective person root-filter and the fisheye person root-filter. The learned spatial model

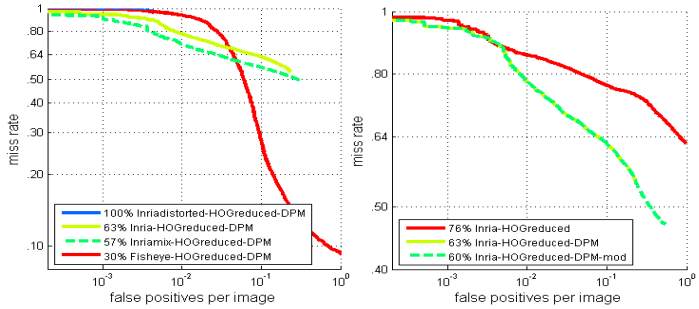

Figure 5: Detection performance of the DPM approach trained with different datasets(left) and the Adaptive DPM-based approach (right).

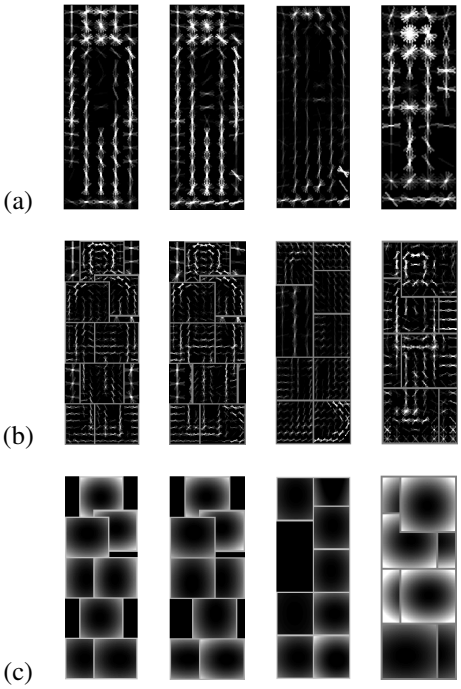

Figure 6: Visualization of DPM models trained with different datasets. From left to right: Inria, Inria-mix, Inria-full-distorted and Fisheye. (a) The coarse root-filter. (b) All the higher resolution part-filters superposed on the root-filter. (c) The spatial model of the location of each part relative to the root.

with anchors position defined by the root filter is thus different. Notice also the correspondence between human body parts and the model parts is not clear anymore.

It is worth mentioning that in the case of the fisheye detector, training and test images are extracted from the same dataset. The images were captured with the same camera and configuration. All these conditions improve the performance of the fisheye detector. Taking into account these remarks, we still believe that the classifier can learn to adapt to radial distortions.

c) Adaptive DPM-based model: Here, instead of enriching the perspective training dataset by simulating fisheye type of distortions, we aim at taking into account the distortion in the detection model. The result in Fig. 5-right show a minor improvement in performance. Our proposed adaptive model show an advantage only on extreme case. In fact, even in the most extreme case, the dislocation of part's anchor $\Delta\tilde{\boldsymbol{v}}_i$ is smaller than 40% of the root filter's size. Meanwhile, the distortion cost is calculated in the local region up to 80% root filter's size around the anchor point. When the response of the part-filter F_i is high, the value will be propagated far enough to take part into the model. Relocating the anchor position is therefore only useful in difficult cases of a weak filter response.

V. Conclusions

Given that the DPM has a very good performance in detecting and recognizing objects on perspective images, we built a fisheye image dataset and evaluated the DPM approach in the context of people detection on fisheye images. It turned out that the deformable models can handle very well the strong radial distortions. We believe that it is possible to build an adaptive DPM-based detector which can solve the problem of object detection in all kinds of non-conventional cameras with known calibration information. The experiments on the proposed method show improvement in performance. One of the drawbacks of all DPM-based approaches is the heavy computational cost. We tent to improve it by combining the camera with a range sensor (Lidar or ultrasonic) to reduce the region of detection on the image. Indeed, this simple combination is helpful in accelerating the detection and reducing false positives in complex texture backgrounds.

References

[1] K.O. Arras, O.M. Mozos, and W. Burgard. Using boosted features for the detection of people in 2d range data. In *IEEE International Conference on Robotics and Automation*, pages 3402–3407. IEEE, 2007.

[2] J.Y. Bouguet. Camera calibration toolbox for matlab. 2004.

[3] M. T. Bui, V. Frémont, D. Boukerroui, and P. Letort. People detection in heavy machines applications. In *IEEE International Conference on Cybernetics and Intelligent Systems*, 2013.

[4] T. Bülow. Spherical diffusion for 3d surface smoothing. *IEEE Trans. Pattern Anal. Machine Intell.*, 2004.

[5] N. Dalal. *Finding people in images and videos*. PhD thesis, 2006.

[6] K. Daniilidis, A. Makadia, and T. Bulow. Image processing in catadioptric planes: Spatiotemporal derivatives and optical flow computation. In *IEEE Workshop on Omnidirectional Vision*, 2002.

[7] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Machine Intell.*, 2011.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 2010.

[10] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. Pattern Anal. Machine Intell.*, 32(9):1627–1645, September 2010.

[11] P. Hansen, P. Corke, and W. Boles. Wide-angle visual feature matching for outdoor localization. *The International Journal of Robotics Research*, 2010.

[12] J. Heikkila and O. Silven. A four-step camera calibration procedure with implicit image correction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1997.

[13] C. Hughes, M. Glavin, E. Jones, and P. Denny. Wide-angle camera technology for automotive applications: a review. *IEEE Trans. Intell. Transport. Syst.*, March 2009.

[14] M. Hussein, F. Porikli, and L. Davis. A comprehensive evaluation framework and a comparative study for human detectors. *IEEE Trans. Intell. Transport. Syst.*, Sep 2009.

[15] M. Lourenço, J.P. Barreto, and F. Vasconcelos. srd-sift: Keypoint detection and matching in images with radial distortion. *IEEE Trans. Robot*, 2012.