LYING-POSE DETECTION WITH TRAINING DATASET EXPANSION

Dao-Xun Xia[1,2,3] *Song-Zhi Su*[1,2*] *Shao-Zi Li*[1,2] *Pierre-Marc Jodoin*[4]

[1] School of Information Science and Technology, Xiamen University, Xiamen, 361005
[2] Fujian Key Laboratory of the Brain-like Intelligent Systems, Xiamen, 361005
[3] School of Mathematics and Computer Science, Guizhou Normal University, Guiyang 550001
[4]Sherbrooke University, Sherbrooke, Québec, Canada

ABSTRACT

We propose a rotation and scale invariant method to locate people lying on the ground. Unlike conventional human-shape detection methods which assume that all human shapes are in upright position, a person lying on the ground can have arbitrary orientation and pose. Accounting for every possible body configuration would thus require a huge training dataset that would be challenging to gather.

In this paper, we propose a method which increases the size of a small training dataset and allows to detect multiple body poses. To do so, our method increases the size of the dataset with a geometric distortion method followed by a rejection sampling method. Then, it automatically identifies K body configurations in the training set, realign it in upright position and trains K SVM classifiers, one for each body configuration. Lying pose detection is then performed by considering a max pooling strategy across all K SVM classifiers.

***Index Terms*—** Lying pose detection, perspective transformation, training set expansion, mean shift.

1. INTRODUCTION

In this paper, we aim at detecting people lying on the ground, an important topic for several applications. One such application is fall detection for elders and disabled people living in smart-homes [1, 2]. In 2012, a report from the World Health Organization revealed that falls are the second leading cause of accidental-injury deaths worldwide and that every year, no less than 37 million falls are severe enough to require medical attention [3]. As a consequence, efficient visual fall detection algorithms is a key element to support elders and disabled to stay home. Lying pose detection can also be used in conjunction with UAVs (unmanned aerial vehicle) for search and rescue missions [4, 5, 6]. With a rising number of UAVs worldwide [7], the need for robust and rotation-invariant object detection methods is becoming a glaring issue.

Unfortunately, only few papers focus on the topic of lying pose detection. In fact, state-of-the-art human-shape detectors mainly focus on pedestrians, rather than people lying on the ground [8, 9, 10]. Although similar to pedestrian detection, lying-pose detection is much more difficult since it cannot assume that human shapes are in upright position. Without this assumption, human shapes can have arbitrary orientation and pose. Furthermore, depending on the camera standpoint, human shapes can suffer from severe perspective distortion (e.g. Fig. 1). Consequently, any common pattern recognition system geared towards lying pose detection would require a large training dataset in order to account for all possible body configuration. However, since people do not normally lay on the ground, such dataset would be challenging to gather. Also, as will be shown in the results section, a part-based human shape detection method such as [11] is far too slow to be useful in practice.

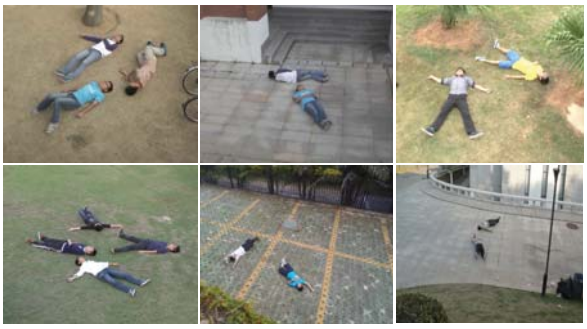

Fig. 1. Example of people lying on the ground with different orientation, pose and perspective distortion.

In this paper, we propose a method that answers the challenges of human lying-pose detection. Also, we exploit the joints information to built a pose-specific classifier, which improves the detection performance, and increase the diversity of the image instance by sample expansion. The method starts with a small annotated training image dataset, *i.e.* a dataset for which every person have been cropped and labeled with a 15-joint skeleton. From there, the number of images in the dataset is increased by applying a series of perspective transformations. This simulates the effect of a moving camera around the persons lying on the ground. Then, a D-dimensional HOG feature is extracted from each body image and K different body poses are localized with K-Means. From these data, a new series of D-dimensional points are generated with a rejection sampling method [12]. Since these newly generated points corresponds to new body poses, they further increases the richness of the dataset. Then, K SVM classifiers are trained, one for each body pose. These K SVM classifiers are then used to locate bodies lying on the ground in new images. This is done following a max pooling criteria.

The paper has the following main contributions:

1. our upsampling method allows to increase the size of the training dataset by accounting for an increased number of geometric distortion and body poses. Results obtained with this upsampling strategy are two times more accurate than without it.
2. By using K SVM classifiers, our method implements a strongly supervised classification procedure whose results are 32% more accurate than for a single SVM;
3. Unlike many other methods, our approach works on single image and does not need a video feed.

2. RELATED WORKS

A limited number of papers have been published on the topic of lying pose detection. In fact, several such methods focus on the more

978-1-4799-5751-4/14/$31.00 ©2014 IEEE

specific problem of fall detection. For example, Wang et al. [2] proposed a deformable part-based model for indoor applications. Given the bounding box of a person lying on the ground, they infer the lymbs of the person and then figure out the pose. Although the method is decently accurate, results reported in the paper have been obtained in a strickly-controlled indoor environment. Mirmahboub et al. [1] proposed a low-cost and easy-to-implement video-based system for human fall detection. With a background subtraction method, they analyse the variation of human silhouette in time and argue that a sudden increase in the size of the silhouette is a strong indication that the person just fell on the ground. That being said, they do not explain how their system accommodates with multiple people and partial occlusion. Toreyin et al. [13] proposed a fall detection system based on the aspect ratio of a bounding box surrounding human shapes. In this method, wavelet coefficients are extracted from the bounding box and then fed to a Hidden Markov Model for classification. Please consider the following survey paper [14] for more details on fall detection methods.

The main inconvenient with fall detection methods is that they need a video feed obtained by a fix camera. These methods are thus inapplicable for videos with a very low frame rate and for applications for which the camera is moving like on a UAV.

As a solution, some authors proposed a single-image lying pose detection method. Andriluka et al. [4] evaluated four state-of-the-art pedestrian detectors, i.e., HOG+SVM [15], deformable part model (DPM) [16], pictorial structure (PS) [17], and poselet based detection [18], in the context of vision-based victim detection from an UAV. The evaluation results show that the two best performing detectors are both built on the pictorial structures framework, and the performance of DPM is better than PS. Also, part-based models are better suited for victim detection than monolithic models [15] and combining visual detectors with inertial sensor data of the UAV will substantially improve the detection performance. That being said, as will be shown in the results section, part-based model can be prohibitively slow [11].

Another appoach is to consider the human lying pose detection problem as a an activity recognition problem. For example, Qian et al. [19] proposed a global feature named *contour coding of motion energy image*, which is combined with local features and a multi-class SVM classifier to recognizing human activities.

3. PROPOSED METHOD

As mentioned previously, the number of body configurations of people lying on the ground can be very large. Bodies can have arbitrary poses and arbitrary perspective distortion due to various camera positions. It is thus very challenging to come out with a complete training dataset which spans across all body configurations. So, instead of creating such a large dataset, we start from a smaller dataset and automatically increase its size with two procedures. The first procedure (which we call *geometric expansion*) increases the number of images in the training dataset by simulating a moving camera around the persons lying on the ground. The second procedure is focused on increasing the number of body poses with the help of a rejection sampling method [12]. The underlying idea is to increase and enrich the dataset on which the SVM classifiers will be trained on.

More specifically, the training stage of our method implements the following four steps : 1) increase the number of images in the training dataset with a geometric expansion method, 2) locate K different body poses from the training images, 3) increase the number of poses with rejection sampling and 4) train K SVM classifiers on the newly expanded dataset.

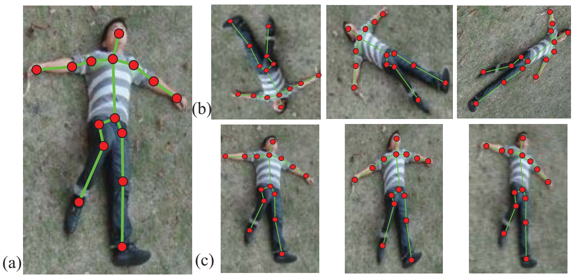

Fig. 2. (a) Annotated picture from our initial dataset. (b) Results obtained after reprojecting the original image into 3 synthetic cameras. (c) shows the images from (b) after being realigned in upright position and normalized.

3.1. Training Step 1: Geometric Expansion

Images from the initial dataset are first manually annotated. That is, every person lying on the ground is cropped and labeled with a 15-joint skeleton. This leads to a series of training images similar to the one in Fig. 2(a). Then, each image is reprojected onto a virtual camera that we move around in order to simulate various perspective distortions.

This camera-reprojection procedure is inspired by Cai *et al.* [20]. According to their method, given a 3D world coordinate system (X_w, Y_w, Z_w), a syntetic camera is positioned at (X_S, Y_S, Z_S) and oriented toward the origin of the world. This camera can later on be moved and reoriented. Then, an input image (like Fig. 2 (a) in our case) is positioned on the XY plane of the world coordinate system. With that configuration, each pixel of the image has a 3D position (x, y, z) with $z = 0$ since the image is on the XY-plane. Each pixel can then be reprojected onto the camera image plane following the projection equation [21]

$$p = K[R|t]w \tag{1}$$

where $p = [x', y', 1]^{\mathrm{T}}$ is a camera pixel in homogeneous coordinates, K is the camera intrinsic matrix, $[R|t]$ the camera rotation-translation extrinsic matrix and $w = [x, y, z = 0, 1]^{\mathrm{T}}$ the image pixel 3D position in homogeneous coordinates.

Cai *et al.* [20] showed that projecting an input image onto a camera plane is equivalent to apply an homography matrix H,

$$H = \begin{bmatrix} -f\cos\kappa & -f\sin\kappa & 0 \\ f\cos\varphi\sin\kappa & -f\cos\varphi\cos\kappa & 0 \\ \sin\varphi\sin\kappa & -\sin\varphi\cos\kappa & -r \end{bmatrix} \tag{2}$$

where f is the camera focal lenth, r the distance between the camera and the origin of the world and (κ, φ) the rotation and elevation angle of the camera. Here, f and r influences the scale of the projected image while κ and φ influences the orientation of the image.

Since the images in the training dataset are all normalized to the same size (roughly 80×160 in our case) only κ and φ need to vary while r and f are set to a constant value. In our case, we take 8 samples between 0° and 80° for κ and 3 samples between 0° and 360° for φ.

Fig. 2 illustrates our geometric expansion procedure. Given a input image (a), the geometric expansion procedure generates a series of perspectively warped images in (b). These new images are then aligned in upright positon and rescaled to a 80×160 size as shown in (c). The images in (c) are then added to the training dataset.

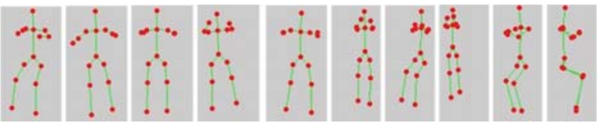

Fig. 3. K-Means centroids for $K = 10$ classes.

Note that this procedure increases by a factor of 24 (8×3) the size of the original training dataset. Also, since perspective transformations have also been applied to the 15-joint skeleton, these newly-generated images all have a 15-joint skeleton.

3.2. Training Step 2 : Lying Pose Clustering

The goal of step 2 is to localize K different body poses out of the M training images. To do so, each skeleton of the training dataset is considered to be a point in a 15×2 dimensional space where 15 stands for the number of joints and 2 is the 2D position of each joint.

Given these M points in this 30 dimensional space, K-means is used to identify K different poses. Fig. 3 shows the resulting centroids of $K = 10$ classes. As can be seen, coherent body poses have been identified such as standing, sitting, and curled position. Let us mention that although similar, the third and fifth poses from the left correspond to people laying on their back and people laying on their front.

3.3. Training Step 3 : Lying Pose Expansion

Now that K different poses have been identified, a D-dimensional HOG feature vector is assigned to each training image. HOG feature vectors are computed as in [15]. Given these HOG feature vectors, the goal of this step is to increase the number of poses on which the SVM classifiers will be trained on. One way of doing so is by increasing the number of points in this D-dimensional HOG feature space. That being said, randomly generating new samples would only add noise to the data. A correct way of increasing the number of samples is by considering the underlying distribution of these data points. In other words, given $Y = \{\vec{y}_1, \vec{y}_2, ..., \vec{y}_n\}$ a set of n HOG feature vectors iid from $P(\vec{y})$, the goal is to generate a new set of HOG samples Y' such that the distribution $P(\vec{y}')$ of the newly generated samples is close to $P(\vec{y})$. We do so with a rejection sampling method [12].

Rejection sampling generates a series of samples iid of a pdf $P(\vec{y})$ given a second pdf $Q(\vec{y})$ that is easier to sample (in our case, $Q(\vec{y})$ is a uniform distribution). A key idea with rejection sampling is that $P(\vec{y}) < MQ(\vec{y})$ where $M > 1$. Given $p(\vec{y})$ and $Q(\vec{y})$, the sampling procedure goes as follows. First, generate a random sample $\vec{y}_i$ iid of $Q(\vec{y})$ as well as a uniform random value $u \in [0, 1]$. If $u < \frac{P(\vec{y}_i)}{MQ(\vec{y}_i}$ then keep $\vec{y}_i$, otherwise reject it. This procedure is repeated up until when the right number of samples has been generated.

Since in our case $P(\vec{y})$ is not known, we estimate it with a Parzen window distribution [22]. We also use mean-shift [23] to find the position $\hat{\vec{y}}$ where $P(\hat{\vec{y}})$ is maximum. In this way, we get to compute M as follows : $M = \frac{P(\hat{\vec{y}})}{1/|\Omega|}$ where Ω is the domain for which the uniform distribution $Q(\vec{y})$ is not zero and $|\Omega|$ its area. With this procedure, we multiply by 3 the number of training HOG samples.

Fig. 4 illustrates in 2D how rejection sampling can be used to increase the size of a training set. Note that the red, green and blue points illustrated classes of 3 different body poses. In this example, our procedure has been executed on each class independantly.

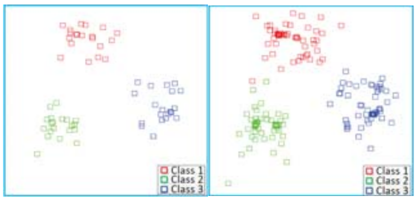

Fig. 4. (Left) scatter plot of 2D feature vectors and (Right) Resulting set of feature vectors after increasing the number of samples with rejection sampling [12].

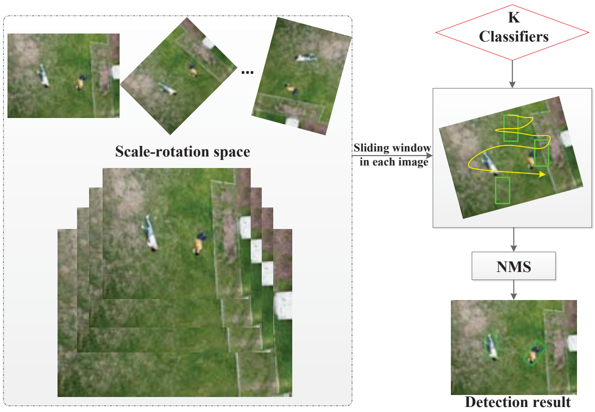

Fig. 5. Sketch map of the lying-pose detection process.

3.4. Training Step 4 : Multi-SVM Training

The last step is to train K SVM classifiers, one for each body pose recovered at step 2. The negative examples used for training are made of non-human images. We use the *libsvm* toolbox to train each SVM model. In order to keep the processing time low, the linear kernel is used.

3.5. Lying-pose Detection

Like most human-shape detection methods, our detection method scans the images with a rectangle window. At each position, a D-dimensional HOG feature vector is extracted (as in training step 3) and fed to all K-SVM classifiers. In order to make the method rotation and scale invariant, this procedure is repeated at different scale and different orientation as shown in Fig. 5. The rotation interval is set to 20 degrees, the initial scale is set to 0.7, the final scale is set to 1.1, and we enlarge the detected images with a scale step of 1.05. The total number of scales is thus $\log(1.1/0.7)/\log(1.05) = 10$.

When a HOG feature fector is fed to an SVM classifier, if the output is greater than a threshold T (0.5 in our case), then the detection result is recorded in a 5-D vector : $(x, y, S, \theta, score)$ where (x, y) is the center of the current window, (S, θ) the scale and rotation angle and $score$ is the SVM output.

Typically, each body lying on the ground generates a series of detection which form a blob in the (x, y, S, θ) space. Once scanning is over, we recover the detected blobs with a mean-shift procedure similar to the one in [24]. To do so, the detected results are normalized in the same scale, namely $(x/S, y/S, \theta)$ and mean-shift is used to find modes. For each mode, we retain the position and orientation with maximum $score$ with a non-maximum suppreson (NMS) procedure. Since the scale of original image is 1.0, the normalized position corresponds to the center of the detected window in the original image.

4. EXPERIMENTS RESULTS

4.1. Dataset

Our lying pose dataset includes indoor and outdoor images. These images were taken inside and outside various buildings, on parking lots, on a beach and in various grassland areas. The cameras were positioned at a height of 2 to 20 meters with different viewing angles and orientation. 30 volunteers participated in this experiment. The dataset contains a total of 1173 images with 0 to 7 persons per image, for a total of 3240 human bodies. In order to gauge performances, we divided the dataset in a training and a testing dataset. As shown in Table 1, the training dataset contains 812 images, 2518 human bodies and 890 negative examples (images void of a human body) while the testing dataset contains 361 images, 722 human bodies and 1293 negative examples.

Table 1. Information on our lying pose dataset.

	Training set		Testing set	
	Positive	*Negative*	*Positive*	*Negative*
Images	812	890	361	1293
Human bodies	2518	-	722	-

4.2. Metrics and Methods

We compared different versions of our method to the HOG-SVM method by Dalal and Triggs [15], the most widely implemented method for human shape detection. Our method is unique on two aspects : it has 2 dataset expansion procedures (geometric expansion and rejection sampling) and its uses K SVM classifiers, one for each body pose. We thus evaluate the impact of these procedures on the results. We tested the use of 1 SVM classifier versus K-SVM classifiers and the use of the original data only (read *Orig* in the result figures) versus the use of geometric expansion (read Orig + GE) and the use of geometric expansion plus rejection sampling (read Orig+GE+RS). Note that our method corresponds to K-SVM with Orig+GE+RS while 1 SVM with Orig corresponds to Dalal and Triggs' method [15] with a rotation in the search space (e.g. Fig. 5).

As in [8], we compared the methods by putting their miss rate against their False positive per window (FPPW) rate in a log-log plot. We also do it for the False positive per image (FPPI) rate. These curves are obtained by varying the detection confidence threshold. FPPW considers the number of true positives against the false positives obtained on images void of human shapes (that is why the testing dataset of Table 1 has negative images). Note that FPPW does not need to perform NMS or other postprocessing. As for FPPI, it considers the total number of detected bounding boxes and compare it with the ground truth bounding boxes. In this case, two bounding boxes match if their overlap exceeds 0.5. Please see [8] for more details.

4.3. Detection Results

Plots in Fig. 6 show that our method (K-SVM Orig+GE+RS) outperforms every other method. For a FPPW rate of 10^{-4}, our method has a miss rate of 17.32% compared to 44% for Dalal and Triggs (SVM Orig). These curves also show that both the use of K-SVM classifiers (instead of one) and our training dataset expansion procedures improve results significantly.

For a FPPW of 10^{-4}, the miss rate of a single SVM method using Orig+GE+RS is of 31%, K-SVM with Orig+GE (so no RS) is 21%, and K-SVM Orig (no GE, no RS) is 25.5%. This shows that both lying pose clustering for K-SVM training, rejection sampling and geometric expansion are effective techniques to improve results. So, according to our experimental validation, our proposed method, especially the geometric expansion and the lying pose clustering, help improve the performance of our detector. When using the FPPI measurement, as shown in the right hand side of Fig. 6, the miss rate of K-SVM Orig+GE+RS is two time lower than that of SVM Orig for an FPPI of 10^{-1}.

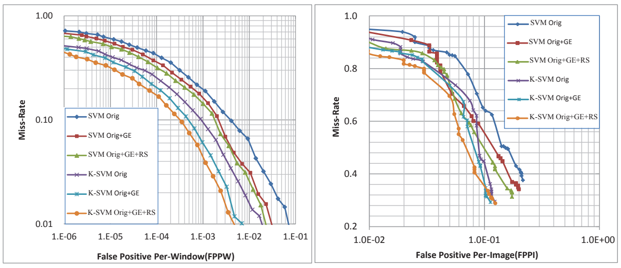

Fig. 6. The performance testing curve. The first column denote False Positive Per Window (FPPW) performance curve. The second column denote False Positive Per Image (FPPI) performance curve.

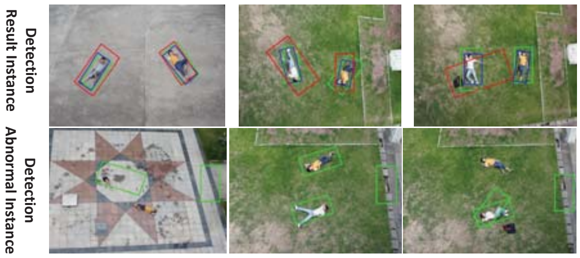

Fig. 7. Blue boxes indicate ground truth, green boxes indicate detection result with our method, and red boxes indicate detection result with Dalal and Triggs' method. Second row shows false positives.

Fig.7 shows detection results. The blue bounding boxes indicate ground truth, the green ones is for our method and the red ones is for Dalal and Triggs (1 SVM + Orig). As one can see, our method is significantly more accurate than the one by Dalal and Triggs. The second row shows example of false detections generated by our method.

Our method takes on average 3 minutes to process a 783x583 image on a personal computer. Note that processing time could be further reduced by using a GPU and various speed-up strategies as in [25]. Let us also mention that method by Yang and Ramanan [11] (with the author's code) take more than 15 minutes per image which prevented us from processing the entire dataset. Early results revealed that their method is 30% less accurate than ours.

5. CONCLUSION

We proposed a method for lying pose detection. Since it is challenging to generate a dataset that includes a large number of body configurations, we propose a method that automatically increases the size of a small training dataset. Our method performs a geometric expansion and then uses rejection sampling to increase the number of HOG feature vectors. Also, in order to account for various body poses, we use K SVM classifiers (instead of one as is usually the case) one for each body pose. Results obtained on 361 images show that the expansion procedure as well as the K-SVM classifiers improve results and outperform method by Dalal and Triggs [15] .

Acknowledgements This work is supported by NSF-China ($No.$61202143), NSF-Fujian ($No.$2013J05100), the Key Projects Fund of Science and Technology in Xiamen ($No.$3502Z20123017), and the Canadian NSERC Discovery Grant 371951.

6. REFERENCES

[1] B. Mirmahboub, S. Samavi, N. Karimi, and S. Shirani, "Automatic monocular system for human fall detection based on variations in silhouette area," *IEEE Trans. Biomed. Eng.*, pp. 427–436, 2013.

[2] S. Wang, S. Zabir, and B. Leibe, "Lying pose recognition for elderly fall detection," in *Procs. of Robotics: Science and Systems*, 2011.

[3] "www.who.int/mediacentre/factsheets/fs344/en/," .

[4] M. Andriluka, P. Schnitzspan, J. Meyer, S.Kohlbrecher, K.Petersen, O.Von Stryk, S. Roth, and B. Schiele, "Vision based victim detection from unmanned aerial vehicles," in *Proc. IEEE Int. Conf. on Int. Robots and Sys.*, 2010.

[5] P. Doherty and P. Rudol, "A uav search and rescue scenario with human body detection and geolocalization," in *Australian Conference on Artificial Intelligence*, 2007, pp. 1–13.

[6] P. Rudol and P. Doherty, "Human body detection and geolocalization for uav search and rescue missions using color and thermal imagery," in *Proc. IEEE Aerospace Conf.*, 2008, vol. 2008.

[7] T.Cox, C. Nagy, M. Skoog, and U. Somers, "Civil uav capability assessment," *NASA report*, 2004.

[8] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 34, no. 4, pp. 743–761, 2012.

[9] D. Geronimo, A. Lopez, A. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 7, pp. 1239 – 1258, 2010.

[10] S.Z. Su, S.Z Li, S.Y.Chen, G.R.Cai, and Y.D. Wu, "A survey on pedestrian detection," *Acta Elect. Sinica*, vol. 40, no. 4, pp. 814–820, 2012.

[11] Y. Yang and D. Ramanan, "Articulated pose estimation using flexible mixtures of parts," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2011.

[12] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, 2009.

[13] B. U. Toreyin, Y. Dedeogl, and A. E. Cetin, "Hmm based falling person detection using both audio and video," in *in proc IEEE Signal Proc. Comm. Apps*, 2006.

[14] M. Mubashir, L Shao, and L. Seed, "A survey on fall detection: Principles and approaches," *Neurocomputing*, vol. 100, no. 16, pp. 144152, 2013.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2005.

[16] P. F. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan., "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.

[17] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Intern. J. Comput. Vis.*, vol. 61, no. 1, pp. 55–79, 2005.

[18] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *Proc. IEEE Int. Conf. Computer Vision*, 2009.

[19] H. Qian, Y. Mao, W. Xiang, and Z.Wang, "Recognition of human activities using svm multi-class classifier," *Pattern Recognit. Lett.*, vol. 31, no. 1, pp. 100–111, 2010.

[20] Guo-Rong Cai, Pierre-Marc Jodoin, Shao-Zi Li, Yun-Dong Wu, Song-Zhi Su, and Zhen-Kun Huang, "Perspective-sift: An efficient tool for low-altitude remote sensing image registration," *Signal Processing*, vol. 93, no. 11, pp. 3088 – 3110, 2013.

[21] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, second edition, 2004.

[22] Bishop C., *Neural Networks for Pattern Recognition*, Oxford University Press, 1996.

[23] Comaniciu D. and Meer P., "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 5, pp. 603–619, 2002.

[24] Bing Shuai, Y. Cheng, S.Z Li, and S.Z Su, "A hierarchical clustering based non-maximum suppression method in pedestrian detection," in *Intelligent Science and Intelligent Data Engineering*, 2012, pp. 201–209.

[25] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Machine Intell.*, pp. 1–11, 2014.