Human Detection by a Small Autonomous Mobile Robot

Kouhei Takemoto*, Shigeru Takano**, Einoshin Suzuki***

*Graduate School of Systems Life Sciences, Kyushu University
819-0395 Fukuoka, Japan
3sl11012m@sls.kyushu-u.ac.jp
**Department of Informatics, ISEE, Kyushu University
819-0395 Fukuoka, Japan
takano@inf.kyushu-u.ac.jp
***Department of Informatics, ISEE, Kyushu University
819-0395 Fukuoka, Japan
suzuki@inf.kyushu-u.ac.jp

Résumé. Nous proposons une méthode utilisant les histogrammes de gradient orienté (HOG) et les séparateurs à vaste marge (SVM) pour la détection de personnes à partir d'images prises depuis un petit robot mobile autonome. Les travaux antérieurs réalisés dans le domaine de la détection d'êtres humains à partir d'images ne peuvent pas être employés pour ce type d'application car ils supposent que les images sont prises à partir d'une position élevée (au moins la hauteur d'un petit enfant) alors que la taille de notre robot n'est que de 15cm. Nous employons à la fois les HOG et les SVM car cette combinaison de méthodes est reconnue comme étant celle ayant le plus de succès pour la détection de personnes. Pour traiter une grande variété de formes humaines, principalement en raison de la distance existant entre les personnes et le robot, nous avons développé une nouvelle méthode de prédiction à deux étapes utilisant deux types de classificateurs SVM qui reposent sur une estimation de la distance. L'estimation est basée sur une proportion de pixels de couleur de peau dans l'image, ce qui nous permet de clairement séparer notre problème de la détection de corps entier et de celle de corps partiel. Les essais réalisés dans un bureau ont montré des résultats prometteurs de notre méthode avec une valeur de F de 0,93.

1 Introduction

Detecting humans is a fundamental skill that must be possessed by a (mobile) robot operating in a populated environment (Schulz et al., 2003; Shiomi et al., 2005; Nakahara et Yamane, 2005). Extracting and managing knowledge used for such a skill is a challenging and rewarding task for researchers in machine learning and data mining. Existing methods often track humans (Schulz et al., 2003; Shiomi et al., 2005), which is a skill beyond human detection, but rely on expensive sensors such as the laser range finder (Schulz et al., 2003; Shiomi et al., 2005). Human detection using only an image sensor, i.e., a camera, exists (Nakahara et Yamane, 2005) but requires to detect the face of a human. More importantly, each of the existing

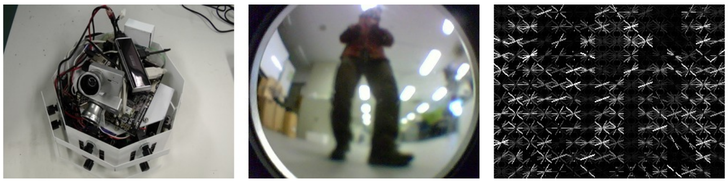

FIG. 1 – *Our robot (left), an image taken by the robot (middle), its HOG representation (right)*

works (Schulz et al., 2003; Shiomi et al., 2005; Nakahara et Yamane, 2005) assumes as the robot a humanoid at least as tall as a human child, i.e., 135cm, which prohibits their usage on an small, autonomous, mobile robot. Such a robot is highly promising to become popular among general consumers due to its small size and low cost, once it is equipped with fundamental skills to operate in a populated environment. However, human detection for such a small robot is challenging because it takes images from a low position, resulting in a variety of shapes of humans to be detected. It should be also noted that using expensive sensors should be avoided to keep the cost of the robot low.

Human detection from one image or a video sequence has known a remarkable progress (Mohan et al., 2001; Dalal et Triggs, 2005; Dalal et al., 2006; Zhu et al., 2006; Iwahori et al., 2010; Lu et al., 2009; Pang et al., 2011). Many of them use image databases of humans taken from a camera either placed on a high position or taken by another human (Mohan et al., 2001; Dalal et Triggs, 2005; Dalal et al., 2006; Zhu et al., 2006; Iwahori et al., 2010; Pang et al., 2011). Lu et al. (2009) uses video sequences of ice hokey games and soccer games which are filmed at a high position. Iwahori et al. (2010) also uses images taken from a camera placed on the ceiling. Apparently none of them may be used for a small robot.

Many of the above works use Histograms of Oriented Gradients (HOG), which are feature descriptors used in object detection (Dalal et Triggs, 2005). HOG has been successfully used in human detection (Dalal et Triggs, 2005; Dalal et al., 2006; Iwahori et al., 2010; Lu et al., 2009; Pang et al., 2011). Pang et al. (2011) states that the combination of HOG and Support Vector Machine (SVM) (Vapnik, 1992) is the most successful human detection algorithm. In this paper, we apply this combination, i.e., HOG + SVM, to human detection by a small, inexpensive, autonomous mobile robot from an image sensor. To cope with a variety of human shapes mainly due to the distance to them, we also propose a two stage prediction method which uses two kinds of SVM classifiers.

2 Problem and our Proposal

2.1 Human Detection by a Small Mobile Robot

Figure 1 left shows the mobile robot used in this work, which is of width 20cm, length 20cm, and height 15cm (Takano et Suzuki, 2011; Boubou et al., 2011). The robot is equipped with two USB cameras, six touch sensors, eight IR sensors, one LED, and two running gears connected to two wheels. The robot has a 1GHz CPU, a 1GB memory unit and a 16GB SDHC card. The robot moves forward and backward, and turns left and right. It navigates autonomously with few collisions to obstacles using the IR sensors and the touch sensors. We use only the upper camera, which takes images of 240×320 pixels in this work.

The robot first navigates in an office and takes one image approximately every one second. It takes in total n images, which are stored in the SD card and then transferred to a PC. The designer labels each image whether it contains a human or not. SVM (Vapnik, 1992) is used to learn one or more classifiers from the images. A binary classifier M, which takes an instance e generated from an image p as its input, is generated from the classifier(s), and then transferred into the memory unit. Finally, the robot moves forward in an office and again takes one image approximately every one second. For each image p, it predicts whether it contains a human based on M and records the result, which is either yes or no, in its memory. As the first step, we assume that a picture contains at most one human.

The problem is challenging due to variations in pose, body shape, appearance, clothing, illumination and background clutter. Moving cameras or backgrounds make it even harder (Dalal et al., 2006). We also define a simplified problem, which replaces the prediction phase by the robot with a cross validation of M on the PC. The results of the both problems are evaluated using recall, precision, and F value.

2.2 Our Solution

A naive, straightforward solution to obtain M for the problem is to use the classifier learnt by SVM as M. Precisely speaking, each image is converted to grayscale and HOG features are extracted to generate one training instance as described later. This method suffers from a huge variety of the body shapes because a human looks highly different when he/she is standing far from or close to the robot. Obviously, the distance h between the robot and the human must be estimated and taken into consideration.

We devised a simple solution based on the ratio $r\%$ of the skin color pixel (Kato et Nakamura, 2005). Kato et Nakamura (2005) states that skin color ranges from 0 to 38 in hue of the HSV color space. Let s be a threshold given by the designer. We split the training images into those $r > s$ and those $r \leq s$, which correspond to images with humans standing far and close, respectively. We apply SVM to the images withe humans standing far and close ot obtain classifiers M_{f} and M_{c}, respectively. As the result, $M = M_{\mathrm{f}}$ if $r > s$ and $M = M_{\mathrm{c}}$ otherwise.

We show the algorithm which generates a feature vector x with HOG from a grayscale image of size $L \times H$ pixels, where $I(x, y)$ represents the brightness of pixel (x, y). It first create HL/c^2 histograms for the corresponding cells with the quadruple for loop, where tranformAngle2Bin($\theta(u, v)$) returns $\lfloor \theta(u, v)/20 \rfloor$ if $\theta(u, v) \leq 180°$ or $\lfloor (\theta(u, v) - 180)/20 \rfloor$ if $\theta(u, v) > 180°$. It then generates x with the rest of the algorithm, where addFeature(x, $H(x_0, y_0), \ldots, H(x_0 + 2c, y_0 + 2c)$) firsts normalizes 9 histograms so that they together form a probability distribution and then determines the $i + 9j + k$th value of x as the kth normalized value of the bin of the jth histogram in $H(x_0, y_0), \ldots, H(x_0 + 2c, y_0 + 2c)$.

Algorithm 1 Generation of a feature vector with HOG

```
INPUT : Grayscale image ($L \times H$ pixels) $I(x, y)$
OUTPUT : feature vector x
for $x_0 = 0$ TO $L - c$ STEP $c$ do
  for $y_0 = 0$ TO $H - c$ STEP $c$ do
    for $u = x_0$ TO $c - 1$ STEP 1 do
      for $v = y_0$ TO $c - 1$ STEP 1 do
        if $0 \leq u \pm 1 \leq L$ AND $0 \leq v \pm 1 \leq H$ then
          $f_{\mathrm{x}}(u, v) = I(u + 1, v) - I(u - 1, v)$.
          $f_{\mathrm{y}}(u, v) = I(u, v + 1) - I(u, v + 1)$.
          $m(u, v) = \sqrt{f_x(x, y)^2 + f_y(x, y)^2}$
          $\theta(u, v) = \tan^{-1} \frac{f_x(x,y)}{f_y(x,y)}$
          Add $m(u, v)$ to the bin tranformAngle2Bin($\theta(u, v)$) of histogram $H(x_0, y_0)$
        end if
      end for
    end for
  end for
end for
i = 1
for $x_0 = 0$ TO $L - 3c$ STEP $3c$ do
  for $y_0 = 0$ TO $H - 3c$ STEP $3c$ do
    addFeature(x, $H(x_0, y_0), \ldots, H(x_0 + 2c, y_0 + 2c)$)
    i = i+81
  end for
end for
RETURN x
```

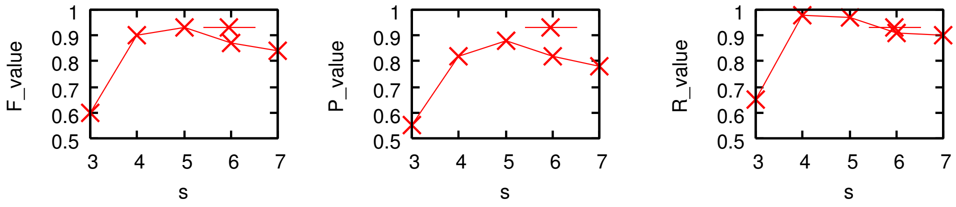

FIG. 2 – *The results of experiments varying s*

3 Experimental Evaluation

3.1 Experiments

All experiments were carried out in the office with three subjects standing closer than 2m, as the whole body gradient is not fully taken otherwise. We use SVMlight (Joachims, 1998) with its default setting without using the kernel function. The PC has one Intel Core i7 processor running at 2.6GHz with 8GB of memory. We use 10-fold cross validation with the experiments on PC. The training data set was generated from 100 images with humans and 100 images without humans and $c = 20$.

First, we tested our proposal on PC by varying s between 3% and 7% with step 1%. Figure 2 shows the results of experiments and we see that $s = 5$ gives the best result in terms of F valut, i.e., F=0.93. We found that $s = 5$ corresponds to approximately $h = 1$m, at which the image contains the whole human body, from the feet to the head. The best performance is achieved probably due to the clear separation of the two cases : whole body detection and partial body detection. Secondly, we tested the naive method with one SVM classifier on PC The accuracy, recall, and precision were 0.85, 0.91, and 0.89, which justifies our proposal of using two SVM classifiers.

Finally, we tested our proposal onboard and the accuracy, recall, and precision were 0.58, 0.59, and 0.56, respectively. As the reason of this drastic drop, we found that the pictures are often misclassified as positive, mainly due to objects which have similar color to skin and were not present in the training images, e.g., a cardboard. As the results on PC are highly encouraging, we think including such images in the training as well as the use of near misses would resolve the problem.

Acknowledgment

A part of this research was supported by Strategic International Cooperative Program funded by Japan Science and Technology Agency (JST). We are grateful to Fabrice Muhlenbach and Philippe Lenca for their help in writing our abstract in French.

Références

Boubou, S., A. Kouno, et E. Suzuki (2011). Implementing Camshift on a Mobile Robot for Person Tracking and Pursuit. In *Proc. Eleventh IEEE International Conference on Data Mining Workshops (ICDMW 2011).* (accepted for publication).

Dalal, N. et B. Triggs (2005). Histograms of Oriented Gradients for Human Detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Volume 1, pp. 886–893.

Dalal, N., B. Triggs, et C. Schmid (2006). Human Detection Using Oriented Histograms of Flow and Appearance. In *Computer Vision–ECCV 2006, LNCS 3952*, pp. 428–441. Springer.

Iwahori, Y., Y. Yamauchi, H. Fujiyoshi, et T. Kanade (2010). People Detection Based on Co-occurrence of Appearance and Spatio-Temporal Features. *National Institute of Informatics Transactions on Progress in Informatics* (7), 33–42.

Joachims, T. (1998). Text Categorization with Support Vector Machines : Learning with Many Relevant Features. In *Machine Learning : ECML-98, LNCS 1398*, pp. 137–142. Springer.

Kato, Y. et O. Nakamura (2005). High-accuracy Extraction of Faces in Consideration of Person with Spectacles. *IEEJ Transactions on Electronics, Information and Systems 125*, 1018–1023.

Lu, W., K. Okuma, et J. Little (2009). Tracking and Recognizing Actions of Multiple Hockey Players Using the Boosted Particle Filter. *Image and Vision Computing 27*(1-2), 189–205.

Mohan, A., C. Papageorgiou, et T. Poggio (2001). Example-Based Object Detection in Images by Components. *IEEE Trans. Pattern Anal. Mach. Intell. 23*(4), 349–361.

Nakahara, T. et T. Yamane (2005). Human Detection Method for Autonomous Mobile Robots. *MEW Technical Report 53*(2), 81–85. (in Japanese).

Pang, Y., Y. Yuan, X. Li, et J. Pan (2011). Efficient HOG Human Detection. *Signal Processing 91*(4), 773–781.

Schulz, D., W. Burgard, D. Fox, et A. Cremers (2003). People Tracking with Mobile Robots using Sample-based Joint Probabilistic Data Association Filters. *The International Journal of Robotics Research 22*(2), 99.

Shiomi, M., T. Miyashita, et H. Ishiguro (2005). Multisensor Based Human Tracking Behaviors with Markov Chain Monte Carlo Algorithms for Active Communication Robots. *Journal of the Robotics Society of Japan 23*(2), 220–228. (in Japanese).

Takano, S. et E. Suzuki (2011). New Object Detection for On-board Robot Vision by Lifting Complex Wavelet Transforms. In *Proc. Eleventh IEEE International Conference on Data Mining Workshops (ICDMW 2011).* (accepted for publication).

Vapnik, V. (1992). Principles of Risk Minimization for Learning Theory. *Advances in neural information processing systems 4*, 831–838.

Zhu, Q., M. Yeh, K. Cheng, et S. Avidan (2006). Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, Volume 2, pp. 1491–1498.

Summary

We propose a human detection method using HOG and SVM from an image by a small autonomous mobile robot. Existing works for human detection from images cannot be used for our purpose because they assume that the images are taken from a high position, at least at the height of a small human child, while our robot is of 15cm height. The combination of HOG and SVM is known as the most successful human detection method so we adopt it. To cope with a wide variety of human shapes mainly due to the distance to them, we devised a two stage prediction method which uses two kinds of SVM classifiers based on an estimation of the distance. The estimation is based on the ratio of the skin color pixel in the image, which allows us to clearly separate our problem into whole body detection and partial body detection. Experiments in an office showed promising results of our method with F value 0.93.