Real-time Human Detection based on Personness Estimation

Kyuwon Kim\textsuperscript{1,2}
q1.kim@samsung.com

Kwanghoon Sohn\textsuperscript{1}
khsohn@yonsei.ac.kr

\textsuperscript{1} School of Electrical and Electronic Engineering
Yonsei University
Republic of Korea
\textsuperscript{2} Mobile Division
Samsung Electronics
Republic of Korea

Even though deformable part model (DPM) [3] is one of the most popular human detection methods [1], it is difficult to see a practical application of DPM human detection on mobile devices due to its computational overhead. On the other hand, many objectness (a.k.a. detection proposals) measure methods have been recently proposed [4]. An objectness measure generates object windows that are likely to contain generic objects, and allows avoiding an exhaustive search. Its intention to improve detection speed seems to be perfectly matched with DPM human detection in mobile devices. However, most objectness measure methods reviewed and evaluated in [4] are not well suited for real-time detection because all objectness measure methods (except for the binarized normed gradients (BING) method [2]) spend more than 250 milliseconds.

For this reason, we propose a more efficient and accurate method for estimating human windows using the normed gradients with color feature in order to enhance detection rate within a short period of time. In this paper, we are interested in humans, not generic objects. So we name objectness estimation for humans simply as personness estimation.

To rapidly determine the priority order of each feature vector of an image with a supervised approach, we make use of the skin color model and the fast NG feature (or simply Bing) proposed by Cheng et al. [2]. An NG feature is a 64 dimensional vector describing the magnitude (a.k.a. norm) of the gradients of an $8 \times 8$ downsampled image. To examine every image window for the objectness measure, a magnitude map of gradients is resized to 36 predefined quantized shapes (or quantized sizes in [2]). The results are called NG maps. By calculating the correlation values between the NG maps and $w \in \mathbb{R}^{8 \times 8}$, we can obtain a filter score for each window. A quantized shape is ignored in the predicting stage if the quantized shape has less than or equal to 50 positive samples in the training stage. The final objectness score is calculated as follows:

$$o_i = v_1 \cdot s_{(i,x,y)} + t_i \tag{1}$$

where $v_1, t_i \in \mathbb{R}$ and $s_{(i,x,y)}$ are learned coefficient, a bias term and the filter score at position $(x, y)$ of each quantized shape $i$. $s_{(i,x,y)}$ is obtained using the first stage linear SVM, and $v_1$ and $t_i$ are obtained using the second stage linear SVM.

Figure 1: NG features for personness and a $8 \times 8$ downsampled example: (a) Personness filter $w_p$ for human detection. (b) The $8 \times 8$ downsampled image. The yellow box emphasizes the head part

Even though SVM training on various object categories makes the original linear model $w \in \mathbb{R}^{8 \times 8}$ work for generic objects, it degrades the performance of the SVM classifier for single-category object detection. Therefore, we generate the new linear model $w_p \in \mathbb{R}^{8 \times 8}$ shown in Fig. 1(a) by training the linear SVM only on humans that our DPM detector can detect. This restricted training technique also reduces the number of quantized shapes to consider. Fig. 1(a) illustrates that $w_p \in \mathbb{R}^{8 \times 8}$ surprisingly places more confidence in the shoulder and head regions. After training the linear SVM on VOC 2007 dataset, we choose four points $(3, 1), (3, 2), (4, 1), (4, 2)$ to extract skin color information. Head and neck are usually located in these four points as shown in the Fig. 1(b). When function $\text{skin}(x, y)$ returns a binary value indicating whether the point $(x, y)$ is skin color according to a skin color model, the skin score at position $(x, y)$ is

$$c_{(i,x,y)} = \frac{1}{2} \sum_{k=1}^{4} \sum_{j=1}^{2} \text{skin}(x+k, y+j) \tag{2}$$

And the new personness score is as follows:

$$p_i = v_1 \cdot s_{(i,x,y)} + u_i \cdot c_{(i,x,y)} + t_i \tag{3}$$

where $u_i \in \mathbb{R}$ is learned coefficient for skin score using the second stage linear SVM.

To make good use of personness estimation, we divide DPM into three stages: setup, detection, and evaluation. In the setup stage, DPM builds a feature pyramid, and initializes three pyramids: filter response pyramid, probe flag pyramid and convolution flag pyramid. In the detection stage, filter responses are computed between root/part filters and some selected feature vectors. With width $w$, height $h$, and center position $(x+w/2, y+h/2)$ of a window, we can pick the corresponding $(l, \bar{x}, \bar{y})$ where $l$ is the scale level of the filter response pyramid, and $(\bar{x}, \bar{y})$ is location of the filter response map to select feature vectors to be convolved. We call the exact response-map coordinates $(l, \bar{x}, \bar{y})$, which correspond to a given window, pinpoints. At this conversion from windows to pinpoints, one should keep in mind that a window for personness was quantized, so it might not be completely fitted to a real human window. Therefore, two important factors are considered for the window-pinpoints conversion: 1) aspect-ratio threshold 2) patch size for pinpoint. After DPM performs convolutions between the chosen feature vectors and a root filter, it calculates part filter responses of feature vectors corresponding to the root positions. When time runs out, the detector finishes computing filter responses, and moves to the evaluation stage. In the evaluation stage, the detector computes the overall root scores as discussed in [3].

In conclusion, we proposed personness estimation that generates promising human windows within a short period of time for real-time human detection. To show the efficiency and practicality of personness estimation, we designed a real-time DPM human detection that makes effective use of personness estimation.