Nguyen TV, Feng J, Yan S. Seeing human weight from a single RGB-D image. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 29(5): 777–784 Sept. 2014. DOI 10.1007/s11390-014-1467-0

Seeing Human Weight from a Single RGB-D Image

Tam V. Nguyen[1], Jiashi Feng[2] (冯佳时), and Shuicheng Yan[3] (颜水成), *Senior Member*, *IEEE*

[1] *ARTIC Centre, Department for Technology, Innovation and Enterprise, Singapore Polytechnic, Singapore 139651, Singapore*

[2] *Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720, U.S.A.*

[3] *Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583, Singapore*

E-mail: vantam@gmail.com; jshfeng@berkeley.edu; eleyans@nus.edu.sg

Received December 25, 2013; revised July 4, 2014.

Abstract Human weight estimation is useful in a variety of potential applications, e.g., targeted advertisement, entertainment scenarios and forensic science. However, estimating weight only from color cues is particularly challenging since these cues are quite sensitive to lighting and imaging conditions. In this article, we propose a novel weight estimator based on a single RGB-D image, which utilizes the visual color cues and depth information. Our main contributions are three-fold. First, we construct the W8-RGBD dataset including RGB-D images of different people with ground truth weight. Second, the novel *sideview shape* feature and the feature fusion model are proposed to facilitate weight estimation. Additionally, we consider gender as another important factor for human weight estimation. Third, we conduct comprehensive experiments using various regression models and feature fusion models on the new weight dataset, and encouraging results are obtained based on the proposed features and models.

Keywords RGB-D image, depth information, human weight estimation

1 Introduction

In many practical situations, such as targeted advertisement and video surveillance, weight information provides useful information for re-identification purposes. For example, in video surveillance, along with some other physical traits (e.g., height, hair color, body-build), the weight is often a part of the description of a person. There have been two different methods adopted to measure the weight of a person: the spring scale and the balance. The former measures the local force of gravity that acts on the person, while the latter is used to compare the weight of the certain person with that of a known standard mass. The usual way to measure human body weight falls into the first case. However, the pervasive implementation of the scale or weight sensor is impractical since it is high-cost and requires the modification of the architecture. Furthermore, it is also inconvenient to request one person to step on the scale. The usage of a scale is also impossible in the scenarios without gravity such as the outer space. Thus, there exists a legitimate need to remotely predict human weight instead of changing the architecture or uncomfortably enforcing a person to use the scale.

Although the specific task of weight estimation from visual cues has not been extensively studied before, there are many related studies analyzing weight patterns. Most weight estimation methods are based on pre-measured body parts[1-3]. Velardo and Dugelay presented a method to estimate the weight of a human body by exploiting anthropometric features, known to be related to human appearance and correlated to the weight[1]. Buckley *et al.* used the pre-measurements of abdominal circumference and thigh circumference from patient files to estimate the patient's weight[2]. However, it is impractical to ask for pre-measurements of human parts in reality.

Depth cameras have been exploited in computer vision for several years, but the high price and the poor quality of such devices have limited their applicability. Recently, with the invention of the low-cost Microsoft Kinect sensor①, there have emerged numerous researches[4-6]. Most of them have concentrated their efforts on the usage of depth images. Weise *et al.* used the depth map for real-time performance-based facial animation[4]. Meanwhile, Shotton *et al.* introduced body part recognition as an intermediate representation for human pose estimation[5]. Sun *et al.*[7-8] have

Regular Paper

This work is partially supported by Singapore Ministry of Education under Research Grant No. MOE2012-TIF-2-G-016, and also partially by the National Natural Science Foundation of China under Grant No. 61328205.

① Kinect, http://www.xbox.com/kinect, Aug. 2014.

adopted the depth information for sign language recognition. Liu *et al.*[9] introduced an interactive system which recommends the dressings according to the attending event. Most recently, Verlado *et al.* conducted research on weighting a person with no gravity[10]. This work requires human calibration, and then the system utilizes the human skeleton extracted by Kinect SDK② to predict the weight. Their work was evaluated on a small database of 15 subjects. Overall, the weight estimation has never been fully explored by the computer vision community. Weight information is still a challenging feature to be visually extracted.

In this article, we propose a novel approach to estimate human weight using a single RGB-D image taken by Kinect. Fig.1 illustrates the proposed weight estimation framework, characterized by its remarkable simplicity: given only one single RGB-D image captured of the subject, it can estimate the weight with small errors. Additional information such as body height and gender can also be simultaneously estimated in the process. Note that the proposed method is learning-based. For training and testing, we construct a new W8-RGBD③ dataset, which contains 300 RGB-D images of humans whose ground truth weights have been manually collected. Our contributions are as follows.

1) To the best of our knowledge, this is the first study on human weight estimation from a single RGB-D image. And we construct the largest dataset called W8-RGBD for this problem.

2) The novel *sideview shape* feature and the feature fusion model are proposed to tackle human weight estimation.

3) We conduct comprehensive experiments to evaluate the proposed RGB-D based method.

2 Dataset Construction

Since there is no prior research about weight estimation based on single RGB-D images, the W8-RGBD dataset, a collection of RGB-D images, is built for weight estimation research. The images are collected from 190 subjects (university students and staffs). One to three images are captured from each person standing upright, with different locations, dressings, and distances to the camera. Consequently, we collect 300 RGB-D images for the dataset (200 and 100 RGB-D images of male and female subjects, respectively). The ground truth weight of a subject is measured by a usual electronic scale at the same time when his/her RGB-D image is taken.

3 Human Body Extraction

Given an image, the first question in estimating human weight is "where the human is". Human detection has been extensively studied in the past few years, based on gradients, such as histograms of oriented gradients (HOG)[11], or keypoints in the image, such as scale-invariant feature transform (SIFT)[12]. However, the information of the background in the resulted bounding boxes is unhelpful and noisy for the weight estimation. In literature, there exist some methods to detect human from RGB-D images[13-14]. Here, we require the fine contour body in order to tackle the problem. Therefore, we apply depth information in human detection and extraction.

3.1 Preprocessing

All input images are acquired using Kinect, i.e., no other hardware such as a laser scanner is required. Kinect provides both depth and color images at 30 frames per second, based on invisible infra-red projection. Due to the different positions of the depth and RGB cameras in Kinect, RGB and depth images are not well-aligned. We perform calibration on both depth and color cameras to find the transformation to align the images in a similar way as [15].

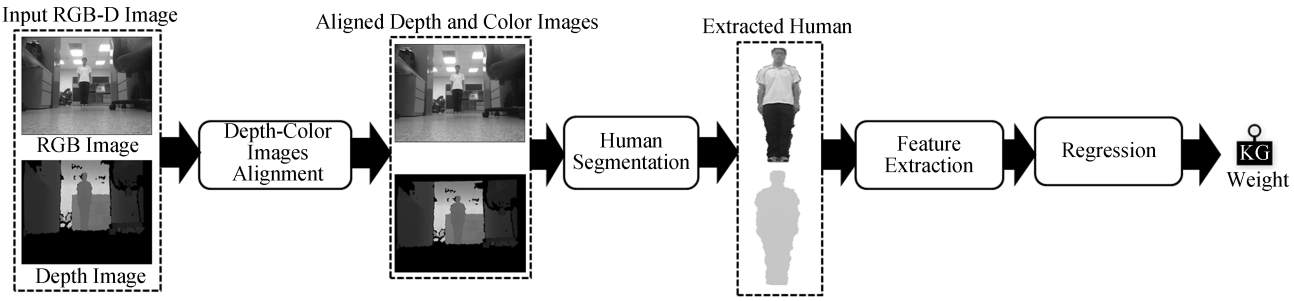


Fig.1. Proposed framework. The depth-color alignment process aligns RGB and depth images. Human detector is applied on the depth image to extract a human. Feature extraction feeds the regression model to estimate human weight.

②OpenNI, http://www.openni.org, Aug. 2014.

③W8 is the abbreviation of "W-Eight" and the dataset consists of color and depth (RGBD) information.

During the image acquisition, it is observed that the quality of the depth image is affected by noises such as the absorb ratio of the surface. Through observation, we find that some parts of human head, hair, and foot may lose depth information. In this work, we propose to smoothen the depth map by filling the loss information as follows. First, the corresponding color image is oversegmented into superpixels using SLIC[16]. Each pixel in the depth image of which the original depth value is equal to 0, i.e., the depth value may be lost due to the noise, is assigned the average depth value of its eight nearest neighbors in the same superpixel. Finally, we apply a conventional Laplacian filter[17] with a 3×3 kernel on pixels whose depth values are equal to 0. Note that the depth image has the same size with that of the color image.

3.2 Human Detection and Segmentation

Detecting and extracting humans in images or videos is a challenging problem due to the variations in pose, clothing, lighting conditions, and complexity of the background. First we apply random sample consensus (RANSAC) on the depth image to remove the floor. Then we parse the RGB-D image into disjoint connected components and perform our human detector.

3.2.1 Searching Connected Components

The conventional method to search the connected components is to perform a flood-fill at every pixel. However, its computational cost is very high. Instead, we adopt the fast finding connected component algorithm④ on the depth image. The idea is to create new blobs and merge them whenever they are discovered to be the same. In particular, the current pixel will be assigned to the same component as its top/left neighbor if their difference is smaller than ε. Here ε is defined as $\frac{\max(D)-\min(D)}{depth\_range}$, where D is a 2D array containing the depth channel of the input RGB-D image, and *depth_range* is empirically set as 21. If both differences are smaller than a threshold ε, the components containing the top and left pixel are merged and then include the current pixel.

3.2.2 Human Extraction

We first run the human detector[11] in the color image. Then we find the overlapping area between the extracted components and the detected human bounding box. If one component has a large overlapping area with a human bounding box ($> 80\%$ in our implementation), we extract its corresponding depth information. All human instances in the W8-RGBD dataset are well extracted by the proposed algorithm. Fig.2 illustrates some of the human segmentation results, compared with the traditional HOG-based human detection method.

4 Features for Human Weight Estimation

To estimate human weight from RGB-D images, we are confronted with two questions, namely, what characteristics are crucial for weight estimation and how to use these characteristics for the task. To answer these questions, we will discuss several feature representation methods together with learning algorithms.

As discussed in [1], human height, width, and gender are distinctive biometrics traits of human beings. Therefore, all of them are considered as human biological features for weight estimation. Additionally, we propose the novel feature, *sideview shape*, for estimating human weight.

4.1 Height Estimation

As human height is an important factor in estimating human weight, we need to compute the real human height in metre unit. Note that one person in photos at different distances to the camera has different heights in pixel. We estimate human height based on the focal length of the depth camera and the depth values of the remaining components. In particular, the component height h can be approximated according to the following geometry equation:

$$h = d\sin(\alpha+\gamma)\Big(\frac{h_a}{h_p + h_{p'}}\Big),$$

Fig.2. Human segmentation samples from W8-RGBD dataset.

④http://xenia.media.mit.edu/∼rahimi/connected, July 2014.

where γ is the angle formed by Kinect motor from its center position, and d is the distance from Kinect to the top of the component. h_a is the height of the human in the depth image, h_p is the distance from the top of the component to the Kinect center and $h_{p'}$ is the distance from the Kinect center to the center of the depth image. h_a, h_p and $h_{p'}$ are in pixel unit. α, the angle between the top of the component and the Kinect center, is computed as $\alpha = \arcsin\frac{p\sin\gamma(h_p+h_{p'})}{h_{p'}d} - \gamma$. The horizontal FOV (field of view) and the vertical FOV are 57.5 and 47.5 degrees, respectively. γ is given by Kinect SDK, and the other parameters are directly computed based upon the depth image. Fig.3(a) illustrates these parameters for the height estimation procedure. The width of the person is analogously computed.

4.2 Human Sideview Shape

The *sideview shape* is constructed from the depth information. Denote $s = (s_1, s_2, \ldots, s_M)$ as the sideview shape of the k-th component in the depth image, M is the number of rows of the k-th component in the depth image, which is calculated based on its *top* and *bottom* values, and s_i is defined by the following equation:

$$s_i = g(D_i), \quad i = 1, 2, \ldots, M,$$

where $D_i = \{D(i,j)|\forall j \ label(i,j) = k\}$, $D(i,j)$ is the depth value at row i and column j, and $g(\cdot)$ is the mean vector function. Note that only the torso or upper body, $s(1 : \lfloor\frac{M}{2}\rfloor)$, is taken into consideration since the lower body may be contaminated by the noises from clothes such as skirt, long-dress, and jeans. We assume that people stand upright. Fig.3(b) demonstrates the *sideview shape* extracted from the upper body.

Eventually, the *sideview shape* feature is linearly rescaled to 200 dimensions.

4.3 Human Gender

We also investigate the usefulness of gender in weight estimation. To this end, we build the regression models for each gender. Here we give a brief introduction about gender estimation and try to include depth information to support such gender estimation.

4.3.1 Traditional Gender Estimation Based on RGB Image

We employ Local Directional Pattern (LDP) features[18] for gender estimation since their descriptors are more robust against lighting variation. LDP provides the same pattern value even with the presence of noises and non-monotonic illumination changes. For training data, FG-NET⑤ and YAMAHA face datasets are utilized. FG-NET and YAMAHA datasets have 1 002 and 8 000 images, respectively. We recapture all images in both datasets with Kinect and then extract faces using a common face detector[19]. All detected faces are processed via histogram equalization and resized to 48×48 pixels. Facial LDP features are ℓ_1-normalized after being extracted from faces. We then apply linear SVM[20] to train the classifier. The accuracy⑥ of our preliminary experiment on 300 RGB images of W8-RGBD dataset is 92%.

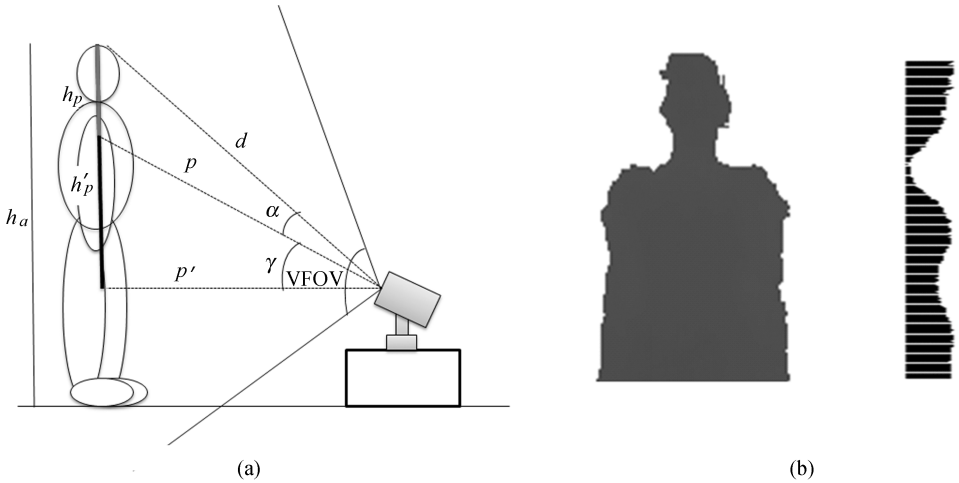


Fig.3. Height estimation and sideview shape illustration. (a) Height estimation based on the distance and angle formed by human head-top. (b) Human sideview shape from human torso scaled to 200-dimension feature. Extracted depth torso (left), and the sideview shape (right).

⑤FG-NET dataset, http://sting.cycollege.ac.cy/∼alanitis/fgnetaging/index.htm, July 2014.

⑥Accuracy is (*TruePositive + TrueNegative*)/*Total*.

4.3.2 Gender Estimation Based on the Sideview Shape

Human faces are not always detectable for the gender estimator in some circumstances (e.g., bad light condition, unusual face pose, and particularly long distance from the camera). Thus, we explore depth information to support gender recognition. We propose miniLBP which is a minimized version of LBP originally developed for texture classification. Instead of computing the LBP code from eight nearest neighbors, miniLBP considers two nearest neighbors of two directions, top and bottom. Given a pixel $c = (x_c, y_c)$, the value of the miniLBP code of c is given by:

$$miniLBP_R(x_c, y_c) = \sum_{p=0}^{2R-1} s(g_p - g_c)2^{2R},$$

where R is the radius of the neighborhood. Here we set $R = 1$, g_c and g_p are the intensities of c and p (a neighbor pixel of c) respectively, and

$$s(x) = \begin{cases} 1, & \text{if } x \geqslant 0, \\ 0, & \text{otherwise.} \end{cases}$$

Depth information here is considered as another cue to classify gender, unlike the traditional methods which depend only on the face. In implementation, we first extract miniLBP features from 200 images which are not included in the W8-RGBD dataset. Then, we use SVM to train the model which will be used for the later experiments. The accuracy of our preliminary experiment on 300 depth images of W8-RGBD dataset is 88%.

5 Human Weight Estimation

5.1 Feature Fusion Models

To estimate human weight, we utilize the features including body height, body width, and aforedefined *sideview shape*. With such features, we aim to answer the second question "how to combine those features altogether". We propose four different feature fusion models (FFM) as follows:

FFM#1: body height, body width, and the sideview-shape are combined into one feature vector;

FFM#2: body area (height × width) and sideview-shape are concatenated as the feature vector;

FFM#3: body height and width only (without sideview-shape feature);

FFM#4: sideview-shape feature only.

5.2 Learning to Predict Human Weight

The input feature extracted from an RGB-D image is considered as a set of vectors $\{\boldsymbol{\beta}_i = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{in})^{\mathrm{T}}\}$. For each input vector, we have a real-valued output human weight value y_i associated with the input. We assume that each output human weight is generated by the input through the following linear equation:

$$f(\boldsymbol{\beta}_i) = \sum_{j=0}^{n} \omega_j \beta_{ij},$$

where f is the estimated weight and w_j is the coefficient (weight) of each feature dimension.

In order to estimate the value of each coefficient ω_j, we minimize the following cost function:

$$\begin{aligned} C(\omega) &= \frac{1}{2}\sum_{i=1}^{m}(y_i - f(\boldsymbol{\beta}_i))^2 \\ &= \frac{1}{2}\sum_{i=1}^{m}(y_i - \sum_{j=0}^{n}\omega_j\beta_{ij})^2. \end{aligned}$$

We consider using two standard regression methods: ℓ_2-regularization and Support Vector Regression (SVR). These methods are simple yet effective and suitable for the problem. The evaluation is conducted on the above two regression methods across four FFMs as mentioned in Subsection 5.1.

In ℓ_2-regularized regression, the closed form is obtained as follows.

$$\begin{aligned} \omega^* &= \arg\min_{\omega} \frac{1}{2}||y - \boldsymbol{\beta}\omega||^2 + \frac{\lambda}{2}||\omega||_{\ell_2} \\ &= (\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\beta}^{\mathrm{T}}y, \end{aligned}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_n)$ is the basis matrix, $\boldsymbol{I}$ denotes the identity matrix, and λ is the regularization parameter.

For Support Vector Regression (SVR), we employ LibSVM[20] with Gaussian Radial Basis Function (RBF) for the kernel function.

6 Experimental Results

We conduct extensive experiments on the W8-RGBD dataset in order to identify the most suitable feature fusion model for human weight estimation. The W8-RGBD dataset contains 200 and 100 RGB-D images of male and female groups respectively. The dataset is randomly split into five batches. Each batch has 40 RGB-D images of male group and 20 RGB-D images of female group. We keep the constraint such that all of the RGB-D images of the same person belong to the same batch. For each experiment, we perform a standard 5-fold cross validation test to evaluate the accuracy of our algorithms on the W8-RGBD dataset. The validation process is repeated five times, with each of the five batches used exactly once as the testing data

and the remaining four batches are used as training data. Each observation is used for testing exactly once. We use the mean absolute error (MAE) to evaluate the accuracy of weight estimation. MAE is widely used in related literature[21-22]. MAE is defined as the average of the absolute errors between the estimated weight and the ground truth weight:

$$MAE = \frac{1}{N_T}\sum_{i=1}^{N_T}|\hat{y_i} - y_i|,$$

where y_i is the ground truth weight for the test image i, $\hat{y_i}$ is the corresponding estimated weight, and N_T is the total number of test images.

In the first experiment, we utilize two regression models which are learned without discriminating the gender. We extract the aforementioned features from 300 instances in the dataset and evaluate various combinations of regression methods and feature fusion models. The MAEs for this experiment on the W8-RGBD dataset are shown in Fig.1. The achieved MAEs of the male are larger than those of the female. Furthermore, ℓ_2-regularized regression yields better results on the minor weight groups. On the contrary, the SVR regression model has the best MAEs on the major weight groups. As can be seen in Table 1, SVR-FFM#2 outperforms other combinations. FFM#2 produces the best results in all regression methods.

In order to examine whether human weight is sensitive to gender, we conduct the second experiment in which regression models are used for each gender individually. Note that, the gender here is the ground truth which is collected manually. We aim to investigate weight estimation MAE (kg) at two gender groups on W8-RGBD with different FFMs. The MAEs for this experiment on the W8-RGBD dataset are shown in Table 2. Similar to the first experiment, the results obtained from SVR are better than those by the ℓ_2 regularization based method.

Another note from the experiment is the effect of FFMs. Unlike the first experiment where FFM#2 dominates, the results in the second experiment show that FFM#2 is effective for female while FFM#1 achieves the lowest MAE for the male in all regression methods. This explicitly indicates that the FFMs we have proposed are beneficial to human weight estimation. Among the combinations, SVR-FFM#2 produces the best results for the female while SVR-FFM#1 produces the best outcomes for the male. Through the results, the sideview-shape feature performs comparably to the human body height and width. The performance of sideview-shape is improved when the individual model is applied for each gender. The best performance is achieved when the sideview-shape feature is combined with human body height and width. The combination is reasonable since the sideview-shape is responsible for the rough "thickness" of the body while the body height and width are utilized to compute the body area. The best results are encouraging with MAE 4.62 kg for the female and 5.59 kg for the male.

The first two experiments demonstrate the individual models for each gender produce better results than the unified one. "Individual model" means that we first detect the gender and then apply the corresponding weight estimator of that gender for the given RGB-D image. Hence, we conduct the third experiment to explore the circumstance in which the ground truth gender is not used. Instead, we use the gender information resulted from our gender classifier. In the case where the face is detected, the gender is classified based on RGB information as discussed in Subsection 4.3.1. Otherwise, the depth information is applied to detect gender as mentioned in Subsection 4.3.2. As learned from the second experiment, we use SVR-FFM#2 for the female group, while SVR-FFM#1 is applied for the male group.

In addition, we also conduct the experiment on human performance, namely human guess in weight estimation. Twelve participants (six males, six females, between 19∼32 years old) are invited to participate in the experiments. Note that none of the twelve participants were involved in W8-RGBD data collection. Each

Table 1. MAEs (kg) of Different Algorithms on W8-RGBD Dataset (Same Model Applied for Both Genders)

Gender	ℓ_2-Regularized Regression				SVR			
(Group)	FFM#1	FFM#2	FFM#3	FFM#4	FFM#1	FFM#2	FFM#3	FFM#4
Female	5.58	5.20	6.59	8.59	5.58	**4.94**	7.25	8.05
Male	6.33	6.30	6.46	9.67	6.16	**6.04**	7.13	9.33

Table 2. MAEs (kg) of Different Algorithms on W8-RGBD Dataset (Individual Model Applied for Each Gender)

Gender	ℓ_2-Regularized Regression				SVR			
(Group)	FFM#1	FFM#2	FFM#3	FFM#4	FFM#1	FFM#2	FFM#3	FFM#4
Female	5.72	5.55	6.13	6.36	5.52	**4.62**	5.88	5.64
Male	6.31	6.69	7.03	7.65	**5.59**	6.67	8.05	7.89

Table 3. MAEs (kg) of Different Models on W8-RGBD Dataset (Listed in Different Gender Groups)

Gender	Unified Model	Individual Models (Groundtruth Gender)	Individual Models (Classified Gender)	Human Performance (on Corresponding Gender)	Human Performance (on Both Genders)
Male	6.04	5.59	5.82	9.17	9.97
Female	4.94	4.62	4.79	5.45	7.06
Total	5.70	5.20	5.41	7.75	8.87

Note: The performance of our framework outperforms the one of human.

participant is shown the images in W8-RGBD dataset in random order and inputs the corresponding estimated human weight. We have two evaluations for human performance. In the first task, we consider all of the estimated weight answered by the participants. In the second task, we only consider the estimated weight answered by the corresponding gender, i.e., the female participants evaluate female instances.

The MAEs of different models are summarized in Table 3 for the male, female, and both, respectively. Even though the results for the individual models based on classified gender do not match the results of the individual models based on ground truth gender, they are slightly better than the unified model's results. It once again shows that the gender is important in estimating human weight and our gender classifier is acceptable for human weight estimation. It also states that the higher accuracy the gender classification achieves, the better the obtained results can be for estimating human weight. Human performance achieves a gain which shows the role of gender in estimating human weight. The superior performance of our method over human shows the advantage of our proposed feature fusion in human weight estimation.

7 Conclusions

To the best of our knowledge, this is the first work on the interesting problem of human weight estimation based on a single RGB-D image. We proposed the new human weight feature and constructed the W8-RGBD dataset. The experiments showed the application of the depth information to estimate human weight is effective. The color cues are useful for correcting depth information and facial gender classification, while the depth information is exploited in order to perform the human segmentation, height estimation, sideview shape feature extraction, and even gender estimation. The depth information deserves to be further investigated to support the existing color information in the future. Our work does not require the calibration and estimates the weight from a single RGB-D image.

In the future, we plan to collect more data with different human poses in order to further investigate how to deal with multiview or occlusion. More advanced estimation or fusion models will be also studied in the future to improve the performance. We also plan to extend our work with dressing recognition and evaluate our work on different range sensors like the laser scanner.

References

[1] Velardo C, Dugelay J L. Weight estimation from visual body appearance. In *Proc. the 4th IEEE International Conference on Biometrics: Theory, Applications and Systems*, Sept. 2010, pp.1-6.

[2] Buckley R G, Stehman C R, DosSantos F L *et al.* Bedside method to estimate actual body weight in the emergency department. *The Journal of Emergency Medicine*, 2012, 42(1): 100-104.

[3] Bloomfield R, Steel E, MacLennan G, Noble D W. Accuracy of weight and height estimation in an intensive care unit: Implications for clinical practice and research. *Critical Care Medicine*, 2006, 34(8): 2153-2157.

[4] Weise T, Bouaziz S, Li H, Pauly M. Realtime performance-based facial animation. *ACM Transactions on Graphics*, 2011, 30(4): Article No.77.

[5] Shotton J, Fitzgibbon A, Cook M *et al.* Real-time human pose recognition in parts from single depth images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2011, pp.1297-1304.

[6] Xia L, Chen C C, Aggarwal J K. Human detection using depth information by Kinect. In *Proc. the 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2011, pp.15-22.

[7] Sun C, Zhang T, Bao B, Xu C, Mei T. Discriminative exemplar coding for sign language recognition with Kinect. *IEEE Transactions on Cybernetics*, 2013, 43(5): 1418-1428.

[8] Sun C, Zhang T, Bao B K, Xu C. Latent support vector machine for sign language recognition with Kinect. In *Proc. the 20th IEEE International Conference on Image Processing*, Sept. 2013, pp.4190-4194.

[9] Liu S, Nguyen T, Feng J *et al.* Hi, magic closet, tell me what to wear! In *Proc. the 20th ACM Multimedia*, Oct.29-Nov.2, 2012, pp.1333-1334.

[10] Velardo C, Dugelay J, Paleari M, Ariano P. Building the space scale or how to weight a person with no gravity. In *Proc. International Conference on Emerging Signal Processing Applications*, Jan. 2012, pp.67-70.

[11] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2005, pp.886-893.

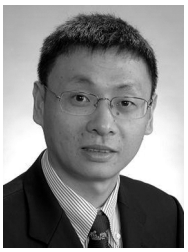
[12] Mikolajczyk K, Schmid C, Zisserman A. Human detection based on a probabilistic assembly of robust part detectors. In *Proc. the 8th European Conference on Computer Vision*, May 2004, pp.69-82.

[13] Basso F, Munaro M, Michieletto S *et al.* Fast and robust multi-people tracking from RGB-D data for a mobile robot. *Advances in Intelligent Systems and Computing*, 2013, 193: 265-276.

[14] Spinello L, Arras K. Leveraging RGB-D data: Adaptive fusion and domain adaptation for object detection. In *Proc. IEEE International Conference on Robotics and Automation*, May 2012, pp.4469-4474.
[15] Janoch A, Karayev S, Jia Y *et al.* A category-level 3-D object dataset: Putting the Kinect to work. In *Proc. IEEE International Conference on Computer Vision Workshops*, Nov. 2011, pp.1168-1174.
[16] Achanta R, Shaji A, Smith K *et al.* SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, 34(11): 2274-2282.
[17] Gonzalez R, Woods R. Digital Image Processing. Addison-Wesley Pub., 1992.
[18] Jabid T, Kabir M H, Chae O. Gender classification using local directional pattern (LDP). In *Proc. the 20th International Conference on Pattern Recognition*, Aug. 2010, pp.2162-2165.
[19] Viola P, Jones M. Robust real-time face detection. *International Journal of Computer Vision*, 2004, 57(2): 137-154.
[20] Chang C C, Lin C J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011, 2(3): Article No.27.
[21] Guo G, Mu G, Fu Y, Huang T S. Human age estimation using bio-inspired features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp.112-119.
[22] Yang M, Zhu S, Lv F, Yu K. Correspondence driven adaptation for human profile recognition. In *Proc. the 24th IEEE Conference on Computer Vision and Pattern Recognition*, June 2011, pp.505-512.

Tam V. Nguyen is a research scientist at ARTIC Centre, Department for Technology, Innovation and Enterprise, Singapore Polytechnic. He obtained his Ph.D. degree in electrical and computer engineering from National University of Singapore in 2013. Prior to that, he obtained his M.E. and B.S. degrees from Chonnam National University, South Korea, in 2009, and University of Science, Vietnam, in 2005, respectively. His research interests include computer vision, multimedia and machine learning. He is the recipient of numerous awards including the Best Paper Award at NUS GSS 2013, the 2nd prize winner of ICPR 2012 Contest on Action Recognition, and the best technical demonstration from ACM MM 2012.

Jiashi Feng is a postdoctoral researcher working with Prof. Trevor Darrell at University of California, Berkeley. He received his Ph.D. degree in electrical and computer engineering from National University of Singapore (NUS) in 2014. He got his bachelor of degree in automation from University of Science and Technology of China (USTC). He is interested in both computer vision and machine learning. In particular, his research work focuses on object recognition, attributes learning, robust optimization, and online and distributed learning.



Shuicheng Yan is currently an associate professor at the Department of Electrical and Computer Engineering at National University of Singapore, and the founding lead of the Learning and Vision Research Group of the department. He received the Ph.D. degree in mathematics from the School of Mathematical Sciences, Peking University, in 2004. Dr. Yan's research areas include machine learning, computer vision and multimedia, and he has authored/coauthored nearly 400 technical papers over a wide range of research topics, with Google Scholar citation $>$ 12 000 times. He is ISI highly-cited researcher 2014, and IAPR Fellow 2014. He has been serving as an associate editor of IEEE TKDE, CVIU, and TCSVT. He received the Best Paper Awards from ACM MM 2013 (Best Paper and Best Student Paper), ACM MM 2012 (Best Demo), PCM 2011, ACM MM 2010, ICME 2010 and ICIMCS 2009, the runner-up prize of ILSVRC 2013, the winner prizes of the classification task in PASCAL VOC 2010∼2012, the winner prize of the segmentation task in PASCAL VOC 2012, the honorable mention prize of the detection task in PASCAL VOC 2010, 2010 TCSVT Best Associate Editor (BAE) Award, 2010 Young Faculty Research Award, 2011 Singapore Young Scientist Award, and 2012 NUS Young Researcher Award.