

Adding Material Embedding to the image2mass Problem

Divya Patel[1], Amar Nath[2], and Rajdeep Niyogi[1(✉)]

[1] Indian Institute of Technology Roorkee, Roorkee 247667, India
divya.patel.iitr@gmail.com, rajdeep.niyogi@cs.iitr.ac.in
[2] SLIET Deemed University, Longowal 148106, India
amarnath@sliet.ac.in

Abstract. An agent has to form a team at run-time in a dynamic environment for some tasks that are not completely specified. For instance, the mass of an object may not be given. Estimating the mass of an object helps in determining the team size, which can then be used for team formation. It has recently been shown that the mass of an object can be estimated from its image. In this paper we augment the existing image2mass model with material embedding. The resulting model has been extensively tested. The experimental results indicate that our model has achieved some improvements on the existing state-of-the-art model for some performance metrics.

Keywords: CNN · image2mass · image2material

1 Introduction

There are some situations (e.g., search and rescue [20], space exploration [1]) where complete knowledge of the world (e.g., description of agents and tasks, total number of agents), is not known in advance. The agents are dispersed in the environment and none of them has a global view. The states and locations of agents may change; agents may enter or exit the environment. The arrival times and locations for tasks are not known a priori. So, an agent who has detected a task needs to form a team at run-time autonomously without the assistance of a central agent. Distributed approaches for team formation in such an environment have been suggested recently [14,15]. In these works, the specification of a task has the following components: name of a task (e.g., move glass container B to location l'), the set of skills/ the number of agents required to execute the task, the location where the task arrived, and the time at which the task arrived.

Another type of task specification has the following components: name of a task (e.g., move glass container B of mass 10kg to location l'), the location where the task arrived, and the time at which the task arrived. The number of agents required to execute the task, however, is not specified. This is because the required number of agents can be determined by (i) the agent who is trying to form a team for the task, and (ii) the physical properties of the object (e.g.,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022 Workshops, LNCS 13377, pp. 77–90, 2022.
https://doi.org/10.1007/978-3-031-10536-4_6

mass, geometry). In this paper, we consider the situation where the mass of the object is not specified. However, the agent can perceive the object through its sensors. Due to the incomplete information, it may appear that the agent cannot figure out the required number of agents for task execution, and thus task execution by the agent fails. One possible approach to come out of the predicament would be to use the past experience of the agent in executing such tasks. Another somewhat easier approach would be to guess/estimate the mass of the object. We take the latter approach.

Our motivation for the latter approach comes from a recent development where the mass of an object can be estimated from its image, referred to as image2mass problem [18]. Thus, given an incomplete task specification, the required number of agents can be found from the estimation of the mass of the object from its image(s). Another development is the image2material model [4,5]. Taking inspirations from such developments, we suggest an augmented model for the image2mass problem that takes into account the material of an object. We show empirically that the performance of our model outperforms that of the existing model.

The remainder of this paper is structured as follows. Related work is given in Sect. 2. The state-of-the art approaches for estimation of mass and material from image is discussed in Sect. 3. In Sect. 4 we suggest a framework for mass estimation. In Sect. 5 we discuss the experimental setup. The experimental results and analysis are given in Sect. 6. Concluding remarks are given in Sect. 7.

2 Related Work

The work [18] first developed a method to estimate the mass of an object from its 2D image. The authors published image2mass dataset: a large dataset of online product information containing images, sizes and mass, and image2mass model (Shape-aware model): a model that estimates the mass from an object's image.

[2] suggest a technique for mass estimation of unknown objects using active thermography. The authors present a method that makes use of active learning. The classification between samples was done using the thermography and proprietary multi-channel neural networks.

Active thermography is a non-contact, remote technology that can investigate the physical attributes of objects in unfamiliar surroundings, such as material classification or thermal characterization. This method involves shining a laser source at the object's surface, examining the dissipation of thermal heat at the object's surface with a thermal camera, and feeding the thermal stream into a machine learning classifier or regressor to determine the object's material class or estimate its corresponding physical properties [2].

The works [3,6,12] discuss the weight estimation for a specific object or material. Their models are tested and trained with same domain data. In [7], mass of two types of foods, i.e., bananas and bread is estimated. All these model are specific and can be used for limited domain.

The works [17,21,22] proposed a methodology that combines image processing and machine learning for measuring the size and mass of several rice kernels,

almonds and pistachio. In all these works mass and size are estimated for a narrow domain like rice, almond, and pistachio. In another work [8], a convolutional neural network (CNN) model was developed to predict the body mass of heifers. This model is also applicable to a specific domain.

3 Mass and Material Estimation from Image

3.1 image2mass: Mass Estimation Using Image

In [18], the authors have addressed the problem of estimating the mass of an object from its image. The work published the image2mass dataset: a large dataset of online product information containing images, sizes and mass. The work suggested the image2mass model: a model that estimates the mass of an object from it's image. Images of ≈3 million products listed with mass and dimensions using the Amazon Marketplace Web Service (MWS) API were collected. A sample of image2mass dataset is shown in Fig. 1 [18], where the dimensions of the images are in inch.

Image	Dimensions	Mass
	(3.75, 1.50, 4.75)	0.1 lbs (0.05 kg)
	(2.88, 2.88, 8.25)	0.35 lbs (0.16 kg)

Fig. 1. Samples from image2mass dataset

image2mass Architecture [18]**:** The deep architecture, referred to as shape-aware model, is used for estimating the mass of an object using RGB image and its dimensions. The central idea is that an object's mass is the product of its density and volume. Density is determined by the object's material and volume by the object's 3D geometry. Both the density and volume information are captured using different neural networks and then the output of these networks are multiplied to estimate the object's mass. In the architecture there are three main modules: geometry module, volume tower, and density tower.

3.2 image2material: Material Recognition Using Image

Recognizing material in real-world image is a challenging task because real-world materials have rich surface texture, geometry, lighting conditions and clutter. The works [4,5] addressed the problem of estimating the material of an object

from its image. A new large-scale material dataset in wild, the Materials in Context Database (MINC) is given in [4,5]. Two approaches for the task of material recognition: (i) to identify the material of a patch of an image, (ii) to segment full images and identify materials in different segments are suggested in [4,5].

Materials in Context Database (MINC) Dataset: This is a collection of images collected from the public, crowd sourced and open surfaces dataset [4]. But this dataset has few limitations such as many categories are not well sampled, and the images come from a single source (Flickr). Therefore, the data from another source of imagery i.e., professional photos on the interior design website Houzz 3 is also gathered. The material annotations using segments, clicks for full scene material recognition, and patches for material classification are gathered.

The MINC-2500 [5] is a subset of MINC for material classification on single patches as shown in Fig. 2. This dataset is balanced with 2500 samples per category, each sample sized 362 × 362. Five random splits of 2250 training and 250 testing samples per category are provided. Each sample comes from a distinct photo (no photo is sampled twice).



Fig. 2. MINC-2500 dataset [5]

Patch Material Classification: In patch material classification as shown in Fig. 3, classifier predicts for the point at the center of the input patch. The remainder of the input patch may be more of the same material or may be contextual information. The classifier was trained to predict with an input patch which is 23.3% of the smaller image dimension when the image is a full photographic composition. [4,5] trained different CNN classifiers like AlexNet [10], ResNet [16], GoogLeNet [19], etc. The best performing CNN architectures GoogLeNet can achieve 85.2% mean class accuracy.

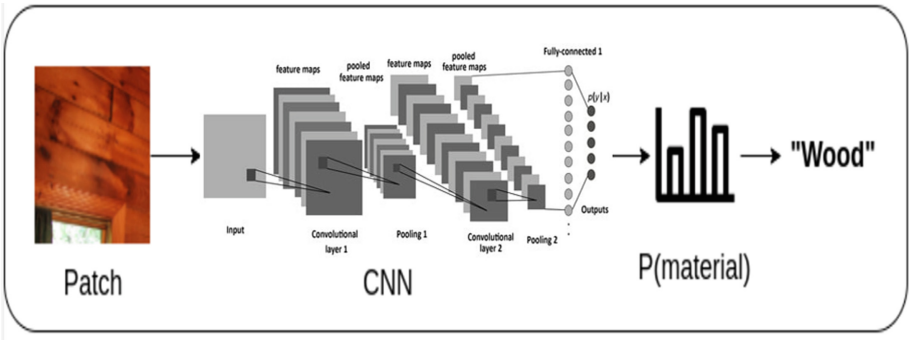

Fig. 3. Architecture of patch material classification [5]

4 Augmented image2mass Model with Material Embedding

In this section we discuss the proposed model obtained by augmenting the state-of-the-art image2mass model with material embedding. In the existing image2mass approach for mass estimation, lack of material information leads to poor performance. So we include explicit material information in the image2mass model. We used the pre-trained image2material model [11,13] to produce a 1024×1 material embedding. Material embedding is an output of average pool layer of the best image2material model (GoogLeNet). It captures the material information of an object in the image. Our model is shown in Fig. 4. We took the estimated mass using the image2mass model, material embedding as input to the fully connected layers, and trained the rest of the network.

5 Experiments

5.1 Material Recognition of image2mass Dataset

We conducted the manual analysis to establish the need of material for estimating mass. An object mass is the product of volume and density. If the volume of the two objects is almost the same, but the object material is different, then the mass will also be different. Different objects have different density. Object with the high density has high mass. Object A is made of metal, and Object B is made of plastic, having the same volume. Metal density is higher than the plastic; therefore, Object A is heavier than Object B. We analyzed the imae2mass dataset for this property. Figure 5 shows that the volume of these two objects is almost the same, but due to different materials their mass are different.

image2material Model: Caffe to PyTorch Conversion:

We used a pre-trained model of image2material to produce the material embedding of objects in the image2mass dataset, using Caffe deep learning

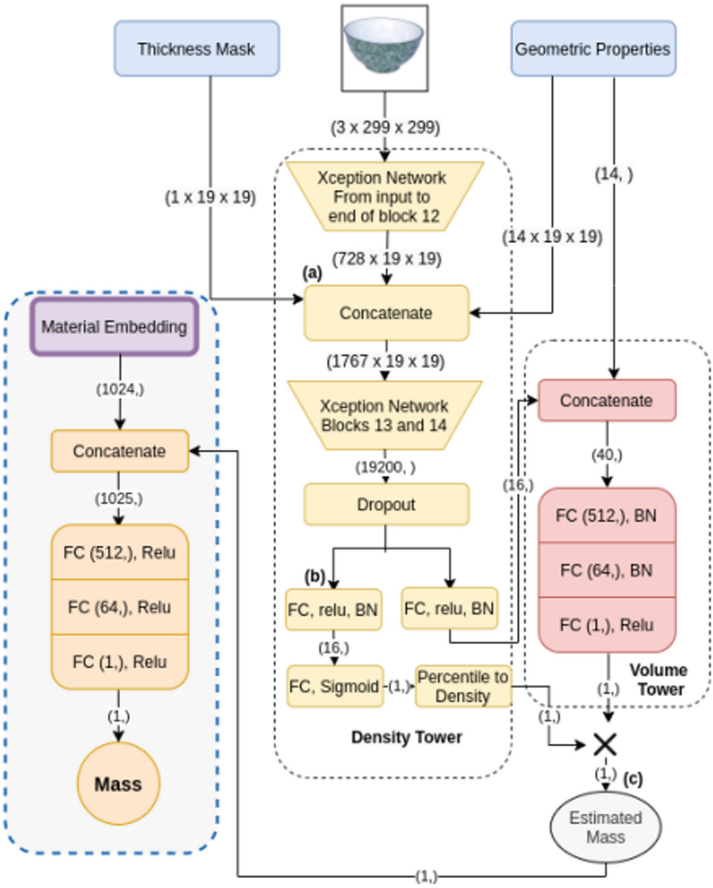

Fig. 4. Augmented shape-aware network model with material embedding



(a) Material label: Plastic, mass: 7.5 lbs, volume: 3672 $inch^3$

(b) Material label: Metal, mass: 11.4 lbs, volume: 3762 $inch^3$

Fig. 5. Relationship between object mass and material

framework. The proposed model is implemented using PyTorch deep learning framework. Therefore, we converted the Caffe model into the PyTorch using Caffe2PyTorch tool[1].

We tested the correctness of the conversion. For an input image (shown in Fig. 6(a)) we estimated the probability of the image belonging to a category using Caffe and PyTorch model. Figure 6(b) shows the output probabilities are almost same. The true label of the input image is 'Brick'. The label predicted is also 'Brick' using Caffe and PyTorch models with probability 0.784312 and 0.796180 respectively.



(a) Input image

Catergory	Caffe Output	PyTorch Output
brick	0.784312	0.796180
carpet	0.000189	0.000177
ceramic	0.000036	0.000034
fabric	0.000444	0.000416
foliage	0.000076	0.000072
food	0.000019	0.000017
glass	0.000766	0.000737
hair	0.000033	0.000032
leather	0.000100	0.000092
metal	0.001428	0.001351
mirror	0.000229	0.000226
other	0.005335	0.004821
painted	0.002624	0.002434
paper	0.000065	0.000057
plastic	0.000600	0.000522
polishedstone	0.009440	0.009143
skin	0.000015	0.000014
sky	0.000008	0.000007
stone	0.046870	0.048022
tile	0.128572	0.119490
wallpaper	0.001159	0.001192
water	0.000028	0.000025
wood	0.017653	0.014945

(b) Probabilities

Fig. 6. Comparison of Caffe and PyTorch results

Analysis of Material Prediction of image2mass Images: image2material model was trained on the MINC dataset, discussed in Sect. 3. For our task, we are using the image2mass dataset for training, validation, and testing. Therefore, we produced material labels of images from the image2mass dataset. We do not have ground truth material labels for the image2mass dataset. To check the correctness of the prediction, we selected random images. We found that very few of the predicted material labels are incorrect. Figure 7 shows the estimation of material on few images of image2mass dataset.

[1] https://github.com/penguinnnnn/Caffe2Pytorch.



Fig. 7. Material recognition of image2Mass

t-SNE Analysis of Material Embedding: We used the average pooling layer output of image2material as material embedding. We do not have the ground truth value of material for the image2mass dataset. Therefore, we did the qualitative analysis of material embedding using t-SNE. Figures 8, 9, and 10 show the t-SNE plot of training, validation, and test set of image2mass dataset. t-SNE plots show that similar material vectors are near to each other and make a cluster. Sometimes material predictions are incorrect but their embeddings are near to the true material embedding. Figure 11 shows that predicted material for the object is sky, whereas the true material label is metal. In t-SNE plot of test dataset, these objects are near to the metal material embedding.

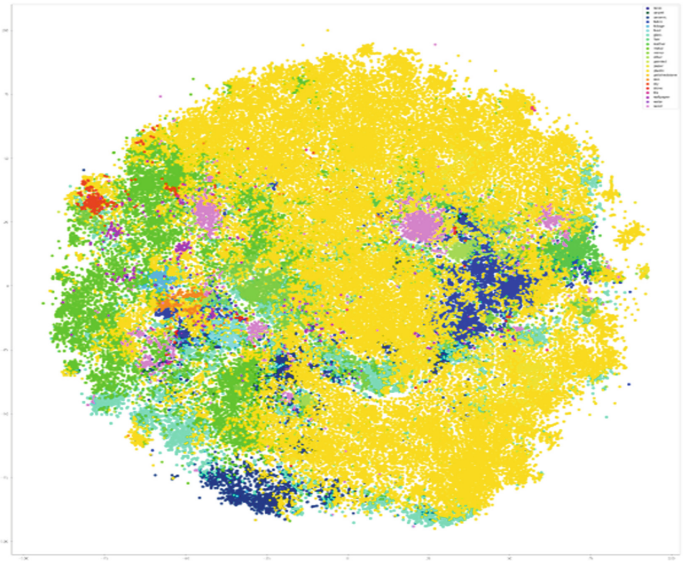

Fig. 8. t-SNE plot of material embedding of the training dataset

5.2 Experimentation Details of Modified Networks

We froze the image2mass model, i.e., weights of that model will not be updated during training, and took the output of the shape-aware model as input to the fully-connected network. The model was trained on an Nvidia DGX-1. We used Pytorch as a deep learning framework. The training was halted after **20** epochs through the data. Validation performance was calculated after each pass, and we used the best model for testing.

A hyper-parameter search was performed. Our model was optimized using Adam [9] with learning rate 0.001. Geometric properties and material embedding were pre-computed for the image2mass dataset. During training, validation and testing, images were randomly rotated by up to 360°, translated by up to 12%, and mirrored 50% of the time.

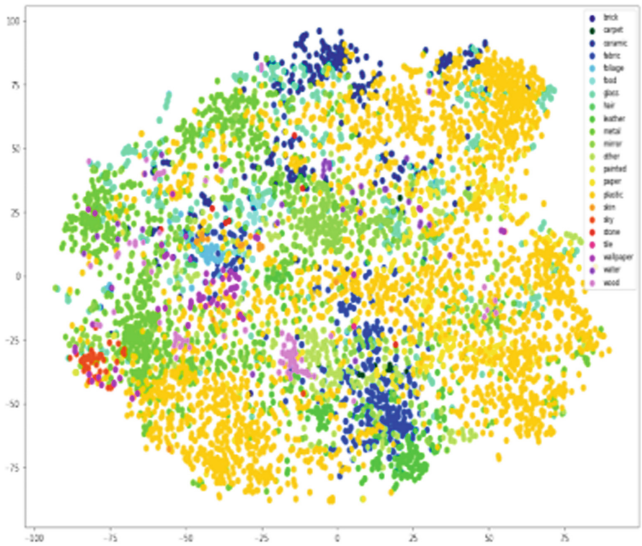

Fig. 9. t-SNE plot of material embedding of the validation dataset

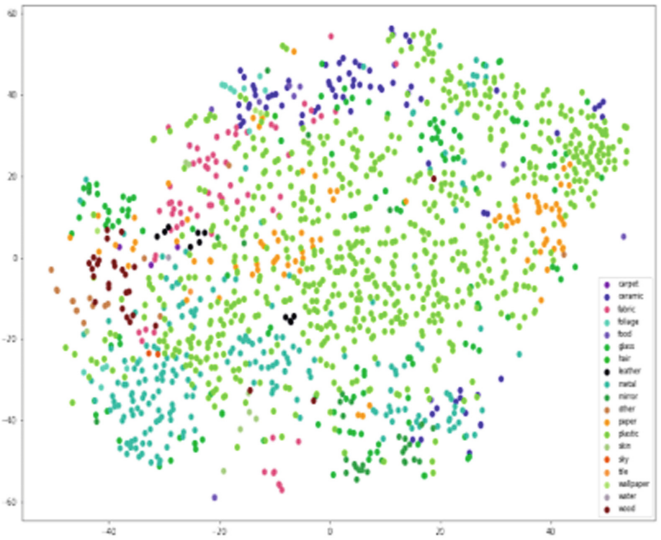

Fig. 10. t-SNE plot of material embedding of the test dataset

6 Results and Analysis

We measured the performance of each model on the basis of the following metrics [18]. Here, the true mass is denoted by t, and the predicted mass by p.



(a) Predicted material: sky
Near to: metal

(b) Predicted material: Sky
Near to: metal

(c) Predicted material: hair
Near to: fabric

Fig. 11. t-SNE analysis on test dataset

- Mean Squared Error (MSE): $\frac{1}{n}\sum_{i=1}^{n}(t_i - p_i)^2$
- Absolute Log Difference Error (ALDE):$|\ln t - \ln p|$, which is our training loss function.
- Absolute Percentage Error (APE): $|\frac{t-p}{t}|$. This metric prefers models that underestimate systematically.

- Coefficient of determination in log space r_{ls}^2 : This metric is a squared correlation between $\ln t$ and $\ln p$.
- Min Ratio Error (MnRE): $min(\frac{p}{t}, \frac{t}{p})$. This is the ideal metric. It has a value of 0 when the model predicts zero or infinity and value 1 when the model makes perfect predictions.
- q: It gives the proportion of the time a model was within a factor of 2 of the true mass.

Figure 12 shows that after adding the material information, we got significant improvement in estimating mass. The three values, for example, 4.95 (**6.23**), [16.91], corresponding to each object in the figure, are to be read as *true value of mass, value obtained using our model, value obtained using the existing image2mass model*; all the values are in *lb*. We considered different types of objects. For instance, when we considered two water bottle containers having similar dimensions (inch) but made of different materials, the estimation of mass by our model is very close to the true value as compared to the image2mass model. This is shown in Fig. 13.

We computed each metric for all models on the image2mass test dataset. Table 1 shows that our model has some improvement over the state-of-the-art image2mass shape-aware model for the metrics MSE, mALDE, and mAPE.

Table 1. Performance of the models on the image2mass test dataset (1435 images)

Model	MSE (↓)	mALDE (↓)	mAPE (↓)	r_{ls}^2	mMnRE (↑)	q (↑)
[18]	2.266 ± 0.230	0.492 ± 0.011	0.689 ± 0.032	0.761	0.662 ± 0.006	0.758 ± 0.011
This paper	1.943 ± 0.179	0.479 ± 0.011	0.616 ± 0.027	0.770	0.669 ± 0.005	0.755 ± 0.011

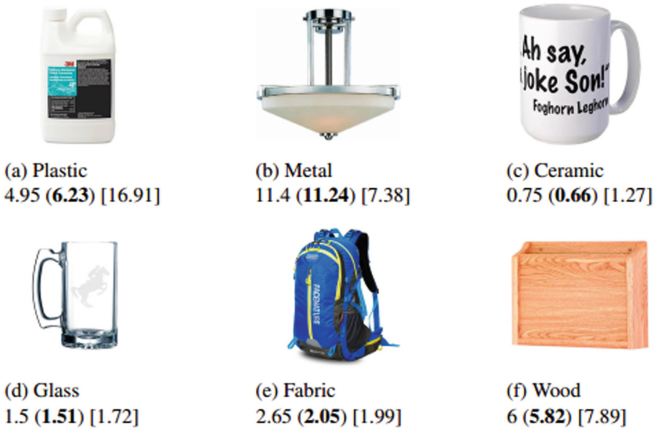

Fig. 12. Improvement after adding material embedding



(a) Material: Metal
Dimensions: 2.75, 2.75, 10.00
Mass: 0.74 (**0.76**) [0.82]

(b) Material: Plastic
Dimensions: 3.25, 4.25, 9.25
Mass: 0.41 (**0.42**) [0.51]

Fig. 13. Mass estimation of water bottles made of different materials

Failure Cases: Figure 14 shows objects for which the estimations made by our model prediction are far from the true values. Many of these errors seem to be due to incorrect material identification. For example, in Fig. 14(a), the correct material of the object is metal, but estimated material is paper, which leads to negative changes in the mass estimation. Figure 14(b) shows that prediction is too high because the estimated material is metal, but correct material of the object is wood. In Fig. 14(d), the prediction is too high because the correct object material is light metal. In Fig. 14(e) the object material is heavy plastic but it is incorrectly identified as plastic and hence the estimated value is too low. This is because the current labels are relatively coarse grained. In other cases, due to lack of knowledge about real-world objects the prediction by our model is not much accurate. For example, in Fig. 14(f) the predicted value is less than the correct value because the object is filled with other components.



(a) Material: Paper
Mass: 13.01 (**6.57**) [13.00]

(b) Material: Metal
Mass: 2.5 (**7.51**) [5.70]

(c) Material: Metal
Mass: 1.6 (**7.24**) [9.49]

(d) Material: Metal
Mass: 0.2 (**0.70**) [0.70]

(e) Material: Plastic
Mass: 10 (**4.61**) [5.01]

(f) Material: Plastic
Mass: 11 (**4.77**) [4.11]

Fig. 14. Failure cases of our model

7 Conclusion

We considered the problem of filling the knowledge gap in incompletely specified tasks. We developed a model, by augmenting the state-of-the-art image2mass model with explicit material information, to estimate the mass of an object from its image. The experimental results indicate that our model has achieved some improvements on the existing state-of-the-art model for some performance metrics.

As part of future work we would like to explore the hyper-parameter space that may increase the likelihood of better estimation. To overcome the problem of the coarse-grained material label, we can categorize material into different abstract classes based on mass (light, medium, heavy) and use these abstract categories to improve material representations. We can use the segmentation of the image into different patches. We can extract the material representation extracted for each patch and aggregate it for utilization.

Acknowledgements. The third author was in part supported by a research grant from Google.

References

1. Agmon, N., Stone, P.: Leading ad hoc agents in joint action settings with multiple teammates. In: AAMAS, pp. 341–348 (2012)
2. Aujeszky, T., Korres, G., Eid, M., Khorrami, F.: Estimating weight of unknown objects using active thermography. Robotics **8**(4), 1–13 (2019)
3. Balaban, M.O., Ünal Şengör, G.F., Soriano, M.G., Ruiz, E.G.: Using image analysis to predict the weight of Alaskan salmon of different species. J. Food Sci. **75**(3), E157–E162 (2010)
4. Bell, S., Upchurch, P., Snavely, N., Bala, K.: Opensurfaces: a richly annotated catalog of surface appearance. ACM Trans. Graph. **32**(4), 1–17 (2013)
5. Bell, S., Upchurch, P., Snavely, N., Bala, K.: Material recognition in the wild with the materials in context database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3479–3487 (2015)
6. Bozkurt, Y., Aktan, S., Ozkaya, S.: Body weight prediction using digital image analysis for slaughtered beef cattle. J. Appl. Anim. Res. **32**(2), 195–198 (2007)
7. Chaithanya, C., Priya, S.: Object weight estimation from 2D images. ARPN J. Eng. Appl. Sci. **10**(17) (2015)
8. Dohmen, R., Catal, C., Liu, Q.: Image-based body mass prediction of heifers using deep neural networks. Biosys. Eng. **204**, 283–293 (2021)
9. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Adv. Neural. Inf. Process. Syst. **25**, 1097–1105 (2012)
11. Krotkov, E.: Robotic perception of material. In: IJCAI, pp. 88–95 (1995)
12. Le, S., Lee, M., Fang, A.: Non-linear image-based regression of body segment parameters. In: Lim, C.T., Goh, J.C.H. (eds.) 13th International Conference on Biomedical Engineering. IFMBE Proceedings, vol. 23, pp. 2038–2042. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-540-92841-6_508

13. Mavrakis, N., Stolkin, R.: Estimation and exploitation of objects' inertial parameters in robotic grasping and manipulation: a survey. Robot. Auton. Syst. **124**, 103374 (2020)
14. Nath, A., AR, A., Niyogi, R.: A distributed approach for road clearance with multi-robot in urban search and rescue environment. Int. J. Intell. Robot. Appl. **3**(4), 392–406 (2019)
15. Nath, A., AR, A., Niyogi, R.: A distributed approach for autonomous cooperative transportation in a dynamic multi-robot environment. In: Proceedings of the 35th Annual ACM Symposium on Applied Computing, pp. 792–799 (2020)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
17. Singh, S.K., Vidyarthi, S.K., Tiwari, R.: Machine learnt image processing to predict weight and size of rice kernels. J. Food Eng. **274**, 109828 (2020)
18. Standley, T., Sener, O., Chen, D., Savarese, S.: image2mass: estimating the mass of an object from its image. In: Conference on Robot Learning, pp. 324–333 (2017)
19. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
20. Tadokoro, S., et al.: The robocup-rescue project: a robotic approach to the disaster mitigation problem. In: Proceedings ICRA, vol. 4, pp. 4089–4094. IEEE (2000)
21. Vidyarthi, S.K., Tiwari, R., Singh, S.K.: Size and mass prediction of almond kernels using machine learning image processing. bioRxiv p. 736348 (2020)
22. Vidyarthi, S.K., Tiwari, R., Singh, S.K., Xiao, H.W.: Prediction of size and mass of pistachio kernels using random forest machine learning. J. Food Process Eng. **43**(9), e13473 (2020)