iWaste: Video-Based Medical Waste Detection and Classification

Junbo Chen[1*], Jeffrey Mao[1*], Cassandra Thiel[2], Yao Wang[1]

***Abstract*— Waste auditing is important for effectively reducing the medical waste generated by resource-intensive operating rooms. To replace the current time-intensive and dangerous manual waste auditing method, we propose a system named iWASTE to detect and classify medical waste based on videos recorded by a camera-equipped waste container. In this pilot study, we collected a video dataset of 4 waste items (gloves, hairnet, mask, and shoecover) and designed a motion detection based preprocessing method to extract and trim useful frames. We propose a novel architecture named R3D+C2D to classify waste videos by combining features learnt by 2D convolutional and 3D convolutional neural networks. The proposed method obtained a promising result (79.99% accuracy) on our challenging dataset.**

***Clinical Relevance*— iWaste enables consistent and effective real-time monitoring of solid waste generation in operating rooms, which can be used to enforce medical waste sorting policies and to identify waste reduction strategies.**

I. INTRODUCTION

A. Clinical Problem: Medical Waste Management

Approximately 1.8 billion kg of waste is produced by healthcare facilities annually [1], one-third of which comes from operating rooms (ORs) [2]. A reliable waste auditing system can help hospitals quantify and characterize waste generation among surgical teams. Data-driven resource scheduling strategies can then be developed to reduce waste. For example, knowing the resource consumption variations among surgical teams can help hospitals standardize clinical practices of low waste generating clinics and target surgeons with high waste generation for intervention. To-date such data collection is conducted manually, in a time-intensive and potentially dangerous manner. To address this challenge, we proposed iWASTE (Intelligent Waste Auditing System for Tracking Emissions), a pilot study designed to replace traditional garbage can in OR; it records entering waste items in video and classifies these items. Such a system enables efficient cataloging of waste generation in surgeries.

B. Waste Classification with Computer Vision

Since AlexNet [3] was introduced, convolutional neural networks have been actively researched in image classification. Recently, neural networks like R2Plus1D [4] proposed different 3D convolutional blocks to solve classification problems in the video domain and achieved state-of-the-art results in action classification benchmarks such as Kinetics [5].

Research on deep learning in waste classification and sorting is very limited compared to problems such as action recognition, which have large benchmark datasets. Currently, there is no prior work done on video classification of medical wastes in a cluttered background. For example, AlexNet and SVM classify only images of plastic, paper and metal [6]. Other benchmark datasets of general waste like VN-trash [7] consists of images on simple backgrounds from multiple classes, including organic trash, inorganic trash, and medical waste. None of the above classify videos in an cluttered background with many similar objects.

C. Challenge in iWASTE

The waste classification methods from Section I-B are not sufficient for real-world application since they classify still images and can not distinguish the newest target item from the background items. To solve this problem, we propose iWASTE, a camera-equipped waste container, which records and classifies video clips of waste as it enters and lands in the waste bin. Our dataset currently includes 4 classes (gloves, hairnet, mask, and shoecover); examples are shown below in Fig. 1. Among this dataset, misclassification of our four classes are equally costly. This is collected as a pilot study to explore the proposed system's viability. We will extend our work to more classes (needle, tissue, etc.) in the future.

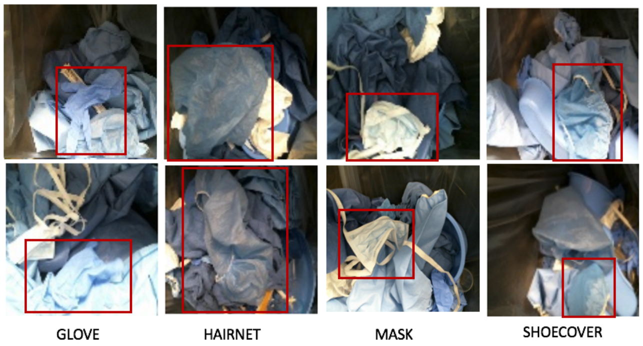

Fig. 1. Two sample frames for each of the four object classes

As shown in Fig. 2a, medical waste falls into the waste container at a fast speed, which leads to blurry images of an object while it is falling. Neural Networks tend to use rich edge and corner information for object classification. This is similar to how a human can tell a glove is a glove by distinguishing the edges of the glove fingers and the corners

[1] Junbo Chen, Jeffrey Mao and Yao Wang are with the Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, Brooklyn, NY 11201, USA. (e-mail: jc7489@nyu.edu, jm7752@nyu.edu, yw523@nyu.edu)

[2] Cassandra Thiel is with the Department of Population Health, Langone School of Medicine, New York University, 227 East 30th Street, New York, NY 10016, USA. (e-mail: Cassandra.Thiel@nyulangone.org)

*Authors contributed equally.

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

where the fingers meet. However, when the glove is falling, its image is often too blurry to correctly identify the object.

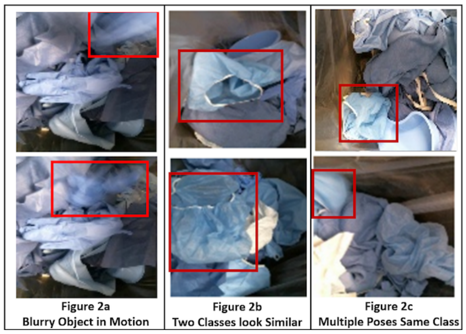

Fig. 2. Various Challenges

Another problem is that many samples of different classes resemble each other. In Fig. 2b, the shoecover on the top is very similar to the hairnet on the bottom in both shape and color. The only subtle differences is texture.

A further challenge is that the same object can fall in multiple poses that do not resemble each other. In Fig. 2c, both objects are shoecovers with little similarity because the shoecover was tossed directly in the top figure, while the shoecover in the bottom figure was scrunched into a ball before disposal. In reality, people are likely to dispose of the same waste item in many different ways.

In comparison, many other common datasets of trash images such as VN-trash [7] have simple backgrounds. This is unrepresentative of the real world as medical waste will be discarded into a cluttered background with many sorts of previously discarded objects, like the backgrounds of Fig. 1 and Fig. 2. For our dataset, the background objects share similar color and are also rich in features. The newest object may also take only a small fraction of the frame, as shown in the bottom image of Fig. 2c. Therefore, a major challenge is to distinguish the newest object added to the pile.

To solve this challenge, we assume one object is thrown into the trash can at a time, and the camera is stationary relative to the can. We further assume that a motion-based segmentation algorithm can automatically recognize the start and end time of each throw. In this project, we focus on classifying the falling object in a short video covering the period from when the object is just thrown until after it lands. Together with the segmentation algorithm, the system can recognize what type of object is thrown in, each time a new object is discarded, thereby automatically cataloging all items thrown into the trash can, until a reset button is hit.

II. Medical Waste Detection and Classification

A. Data Collection

To demonstrate the viability of iWASTE in this pilot study, we collected our dataset in lab with a camera-equipped OR waste bin. Medical waste falling into a waste bin is captured as short videos using a standard metal framed OR linens waste container equipped with web-camera clipped on the rim of the metal frame. The camera is positioned to look down so that it can see the bottom of the plastic bag hanging around the metal frame, as shown in Fig. 3. Four classes of medical waste are collected: shoecover, gloves, hairnet and masks, which are common medical wastes in the OR. We choose these four items for our pilot study in part because they are amorphous and similar in appearance and color. A success in solving this restricted but challenging classification problem will bode well for developing a future system that has to separate many more waste categories.

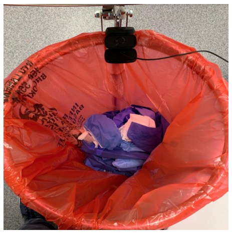

Fig. 3. Trash Can Set-up

We collected a total of 970 videos: 220 Glove videos, 250 hairnet videos, 200 mask videos, and 300 shoecover videos. Each video is 5 second long and has a frame rate of 24 FPS. For each class, 60 videos have spatial resolution of 1920x1080 collected in an earlier set up; the rest have 640x480 resolution, collected in a later setup with a different waste bin and camera. We used the videos collected in the later phase as the training data, and those in the earlier phase for testing, in order to ensure that the trained model can generalize well to very different settings. Each video starts just before a waste item is thrown in, and stops seconds after the object lands. To increase the diversity of our dataset, medical waste was thrown in various poses. The background objects are shuffled for every video i.e. some objects are removed or added and their positions are moved around.

B. Motion-Detection-Based Preprocessing

In our dataset, objects fall at a very fast speed taking only 5-10 frames to land out of a 120 frame video. The frames before the falling object enters contain no useful information. Once the object hits the bottom of the container and becomes stationary, all subsequent frames are almost identical and do not provide additional information. Therefore, we applied motion detection to find and extract frames containing the object's falling process to remove the irrelevant or redundant frames. We found removing these frames greatly improves the accuracy of our neural network, while reducing the computation and memory cost both for model training and for classification using the trained model.

Adaptive GMM (Gaussian Mixture Model) background modeling [8] is applied to detect the motion of falling waste, which classifies pixels as either foreground (moving object) or background. Basic morphological operations including erosion and dilation are applied to post-process the mask generated by GMM for denoising shown in Fig. 4.

With the mask provided by the GMM, the consecutive 16 frames containing the most foreground pixels are extracted as

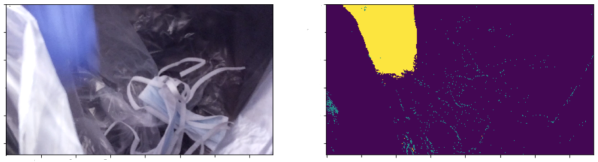

Fig. 4. Motion detection by GMM. Left: glove falling into waste container; Right: moving region identified by GMM

keyframes and are used for classification. Though the GMM background subtraction is not robust against illumination change and background motion in individual frames, the 16-frame window containing the most foreground pixel number can robustly extract the falling process of different waste objects with diverse shapes, poses and background objects.

Aside from removing redundancy in the temporal dimension, we also use the motion detected by GMM to remove the spatial redundancy. As mentioned in Section I, medical waste lacks definable feature when it is falling, and objects may only occupy a small proportion of each frame. Therefore, we applied an algorithm to remove most of the background of each keyframe. For each key frame, a bounding box is generated for the largest cluster of moving pixels. The union of the bounding box area among each frame are taken to crop each of the key frames. Finally, the frames are reshaped to 112x112 for our dataset as seen in Fig. 5 below.

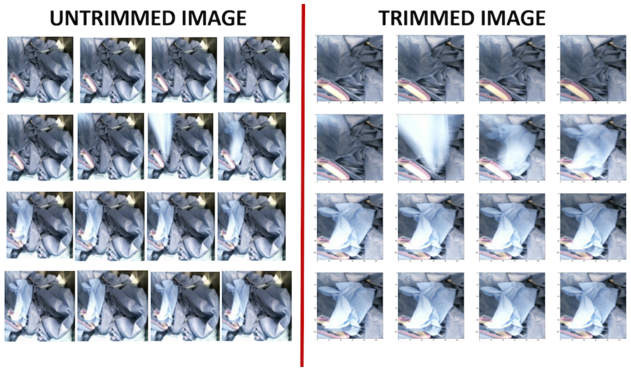

Fig. 5. Spatial Trimming result

C. Deep Learning Model

In our project, we propose a new network architecture, called R3D+C2D network, to classify medical waste based on video. R3D [4] is used as the backbone due to its high performance in 16-frame clips action recognition tasks. Inspired by Tracknet [9], a 2D convolutional network branch is added to R3D to extract useful spatial information for object recognition, leading to the R3D+C2D network.

R3D [4] was proposed by applying residual block to 3D convolution. The architecture of R3D applied in our paper is shown in Fig. 6, which has 13 convolutional layers. The global spatiotemporal pooling converts the convolutional layer's feature maps to a 256 dimensional feature vector.

For R3D+C2D, we add a 2D convolution network branch to R3D. We name this branch C2D, shown in Fig. 6. MobileNetV2 [10] is applied in our C2D branch due to its light weight, which is preferred for resource constrained environments such as smart waste bin. Input video clip is temporally downsampled by 2 and each of the remaining frames is mapped into 32 key features by MobileNetV2, which shares weights across frames.

32 key features from each of the 8 frames give a total of 256 2D spatial features, which are concatenated with the 256 3D features extracted from R3D. These 512 features are placed through a fully connected layer with a softmax activation function at the end for classification as shown in Fig. 6.

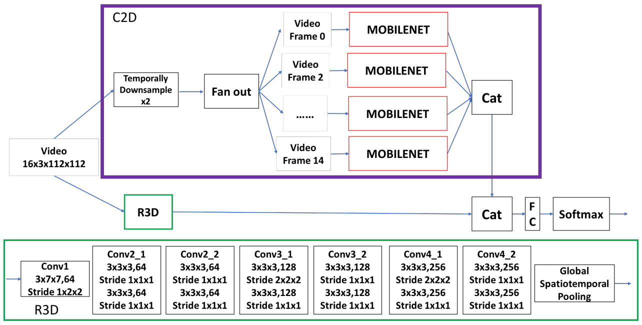

Fig. 6. R3D+C2D architecture. The residual connections in R3D [4] branch are omitted for better readability.

III. Experiments and Result

A. Training Details

Our neural network takes the 16-keyframes of the video clip extracted by motion-detection based preprocessing as an input. Each frame is reshaped to 112x112, making each video a 3D tensor with size 16 (frames) x 3 (RGB) x 112 x 112. For each of the four classes, 60 videos with 1920x1080 original resolution are used as the test set, and the remaining videos with 640x480 original resolution are used as the training set. Since these two datasets are collected with different backgrounds and cameras by different data collectors, good performance on the test set would indicate good generalization capability of the trained model. On training, clips are horizontally flipped with 50% probability. The MobileNetV2 was pretrained on the ImageNet [3].

The SGD (stochastic gradient descent) optimizer with momentum 0.9 and weight decay 0.0005 is applied with an initial learning rate of 0.01. The learning rate gets divided by 10 every 10 epochs while training. The total epoch number is 100, at which time the training loss has converged.

B. Comparison of Results

In order to examine the contribution of R3D and C2D branch, respectively, and the impact of motion-based preprocessing, we trained 6 different models, with results shown in Table I. As expected, R3D+C2D on the spatially trimmed video achieved highest accuracy of 79.99%. When using only the R3D branch, the accuracy is reduced to 57.24%, indicating that the C2D branch extracts additional useful spatial features. However, C2D alone does not provide good test accuracy. This supports the need for a 3D convolutional branch to extract spatiotemporal features from the object's motion. When the network acts on the untrimmed frames which are temporally the same frames though spatially different, the performance is reduced substantially in both the R3D+C2D model and the C2D model. This suggests that removing stationary regions helps the network to focus on the falling object. Since the test set has significant differences from the training set such as very different background and different data collector, the test accuracy shows the good generalization capability and viability of this pilot study.

TABLE I
COMPARISON OF ACCURACY

Network	Test Accuracy, Untrim	Test Accuracy, Trim
R3D+C2D	51.72%	79.99%
R3D	68.96%	57.24%
C2D	43.96%	67.27%

To demonstrate R3D+C2D has learned meaningful features, we applied guided backpropagation [11] on sample videos to generate the saliency maps for the input video frames. Fig. 7 and Fig. 8 show the saliency maps derived for the R3D branch and C2D branch, respectively. The video is a falling mask with another mask and hairnet as the background. The R channel of RGB denotes the focal area.

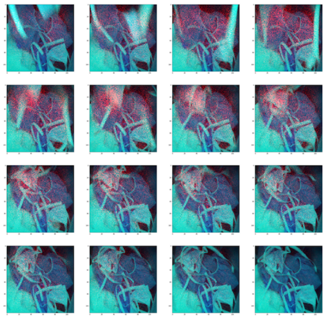

Fig. 7. Saliency map generated from the R3D branch of the trained R3D+C2D model, red dot denotes pixels of high interest

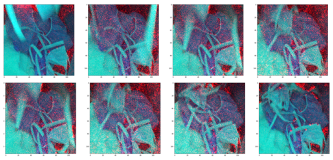

Fig. 8. Saliency map generated from the C2D branch of the trained R3D+C2D model

From the saliency maps, we can see 3D convolutional branch of R3D+C2D seems to track the falling mask as it falls down, and the additional 2D branch, on the other hand, seems to focus on the major outlines of all the objects in the image. Overall, these saliency maps suggest that the network has learnt to track and identify falling objects.

IV. CONCLUSIONS

In this project, a video dataset of 4 medical waste classes was gathered with a variety of backgrounds. A new network architecture, R3D+C2D network, was proposed which in this use case outperformed R3D by a significant margin at 79.99% test accuracy with spatial trimming. This led us to two main conclusions: 1) motion-based preprocessing can help the network to focus on the moving object, which may be particularly important when the training data is limited; and 2) both 3D and 2D convolutions are necessary to provide motion and appearance features necessary to discriminate among similar falling objects.

ACKNOWLEDGMENT

We would like to thank Ruhana Amin and Ava Zheng for collecting the initial 240 video samples. Also, thanks you, Ziming Qiu for your advice on the Saliency Map, and Amirhossein Khalilian-Gourtani for testing RPCA as a GMM alternative. Additional thanks to Abdur 'Mani' Rehman and Jagan Subramanian for early prototypes of iWASTE.

REFERENCES

[1] D. de SA, K. Stephens, M. Kuang, N. Simunovic, J. Karlsson, and O. R. Ayeni, "The direct environmental impact of hip arthroscopy for femoroacetabular impingement: a surgical waste audit of five cases," *Journal of hip preservation surgery*, vol. 3, no. 2, pp. 132–137, 2016.

[2] F. McGain, D. Story, and S. Hendel, "An audit of intensive care unit recyclable waste," *Anaesthesia*, vol. 64, no. 12, pp. 1299–1302, 2009.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[4] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[5] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[6] G. E. Sakr, M. Mokbel, A. Darwich, M. N. Khneisser, and A. Hadi, "Comparing deep learning and support vector machines for autonomous waste sorting," in *2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET)*. IEEE, 2016, pp. 207–212.

[7] A. H. Vo, M. T. Vo, T. Le *et al.*, "A novel framework for trash classification using deep transfer learning," *IEEE Access*, vol. 7, pp. 178 631–178 639, 2019.

[8] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 2. IEEE, 2004, pp. 28–31.

[9] C. Li, G. Dobler, X. Feng, and Y. Wang, "Tracknet: Simultaneous object detection and tracking and its application in traffic video analysis," *arXiv preprint arXiv:1902.01466*, 2019.

[10] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[11] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.