Cloth Material Recognition from Full and Frontal View

YoungJoong Kwon[1]

I. INTRODUCTION

Understanding image and video leads to better reconstruction of the physical world. Previous works focus largely on geometry and visual appearance of the reconstructed scene. In this work, we present a method to recover the material label of cloth from a video. Previous methods on cloth material recovery often require markers or complex experimental set-up to acquire physical properties, or are limited to certain types of images or videos. Our approach leverages the appearance changes of the moving cloth to infer its physical properties. To extract information about the cloth, our method characterizes both the motion and the visual appearance of the cloth geometry. We apply the Residual Network (ResNet) and the Long Short Term Memory (LSTM) neural network to material recovery of cloth from videos.

II. OVERVIEW OF OUR METHOD

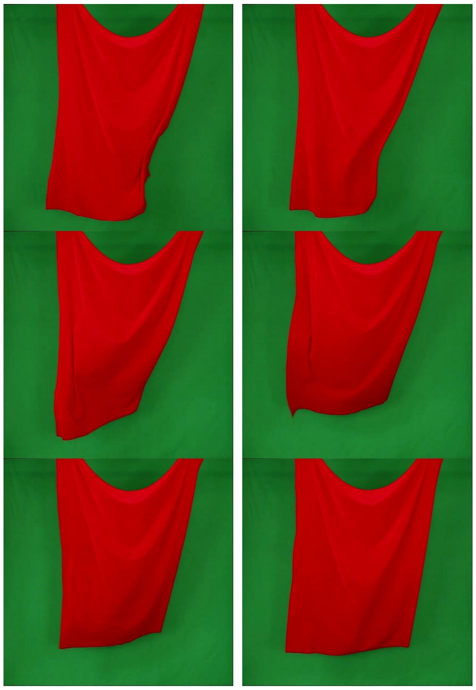
Our cloth material recovery method (Fig. 1) learns an appearance-to-material mapping model from a set of training samples. With the learned mapping model, we perform material-type prediction among 30 classes of material given a recorded video of cloth motion.

III. DATASET

We processed MIT Real Fabric Dataset to make our own dataset.

A. MIT Real Fabric Dataset

MIT Real Fabric Dataset (Fig. 2) consists of 2-3 minuates 90 videos of the 30 different fabrics moving in response to each of the 3 different strengths of wind force. The fabrics span a variety of stiffness and area weights. The material labels include Lycra, Faux Fur, Silk, Silk, Cotton, Wool, Taffeta, Linen, Corduroy, Cotton, Velvet, Fleece, Denim, Upholstery, Upholstery, Pleather, Minky, Damask Upholstry, Flannel Backed Vinyl, Upholstry, Outdoor Polyester, Silk, Wool, Canvas, Nylon Rip Stop, Terry Knit, Lycra, Laminated Cotton, Lycra, Upholstry.



B. Our Dataset

We split each video (29fps) into 5 second-long big segments. Then from each big segments, we subsampled small video segments in rates of 3fps, which leads to length of $5s \times 3fps = 15frames$. So we get $\lfloor \frac{29fps}{3fps} \rfloor = 9$ small videos from each big segments. Therefore, for each video (assume video length is 170 seconds) we get $\lfloor \frac{170s}{5s} \rfloor \times \lfloor \frac{29fps}{3fps} \rfloor = 306$ videos. And each material has 3 videos, so we get about $3 \times 300 = 900$ small videos per material and $30 \times 900 = 27,000$ small videos in total. The exact size of our dataset is 26,136. We used a training, validation, and testing split of 70% / 5% / 25% , respectively.

IV. LEARNING METHOD

In this section, we explain how to establish the mapping between the visual appearance of a moving cloth and its material label using deep neural network.

A. Design Rationale

We propose to combine ResNet with LSTM for our appearance-to-material learning. CNN is used to extract both low- and high-level visual features. LSTM part of the network focuses on temporal motion pattern learning. In the following section, we will briefly introduce our network structure.

[1] YoungJoong Kwon is PhD student of Computer Science, University of North Carolina at Chapel Hill `youngjoong@cs.unc.edu`

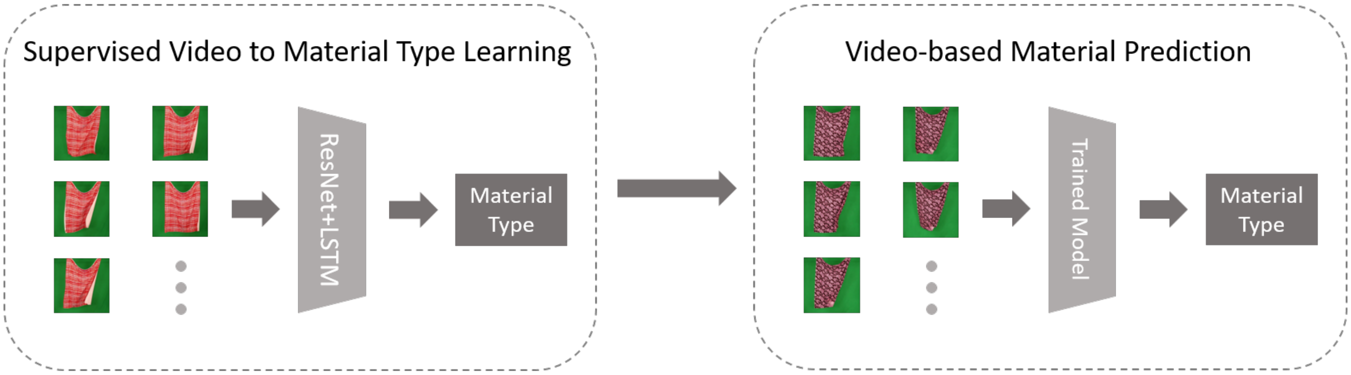

Supervised Video to Material Type Learning
ResNet+LSTM
Material Type
Video-based Material Prediction
Trained Model
Material Type

Fig. 1. Overview of Our Method

B. Residual Network for Hierarchical Visual Feature Learning

The basis of the convolutional operation serves as a filtering operation on an image. Layers of convolutional neural network (CNN) with convolutional kernels of different dimensions extract features at various levels of details.

We applied a ResNet for its ability in hierarchical visual feature selection. However, We removed the fully connected layer of the last stage and directly feed the output of the average pooling layer into each LSTM cell.

C. Recurrent Neural Network for Sequential Pattern Learning

A single image contains a limited amount of information concerning the physics properties of a piece of cloth. But a video can be more powerful to demonstrate how the physics properties, such as the material peoperties of a piece of cloth, can affect its motions. To approximate this mapping between the material properties of the cloth and its sequential movement, we apply the recurrent neural network. Unlike the feed-forward neural network, the recurrent neural network has a feedback loop. The loop connects the output of the current cell to the input of the cell at the next step. The feedback loop act as the ”memory” of the recurrent neural network. With the ”memory”, the recurrent neural network has the ability to gradually extract the pattern of the input sequence.

Following the intuition behind the recurrent neural network, we choose the LSTM instead of the traditional recurrent neural network architecture for its ability to deal with vanishing/exploding gradient and fast convergence to learn the pattern in temporal sequence of data.

D. Network Structure

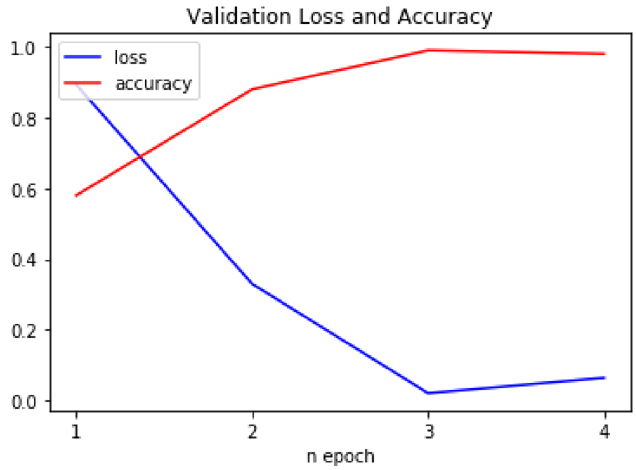
1) ReNet Model: We used Resnet-18 for our work.
2) Input to ResNet: Our input is 4-dimensional tensor which is $timestep \times channel \times height \times width$. Each video consists of 15 frames. Each frame consists of 3 channel (R,G,B) and size of each frame is (256, 256). This leads to input tensor of size (15, 3, 256, 256). We used batch size of 16.
3) Output of ResNet: Feature vector with length of 512 elements for each timestep. Therefore, total size of 15×512 vector.
4) Input to LSTM: 15×512 feature vector.
5) Output of LSTM: vector with length of 30 elements. This vector contains classification score among 30 material classes.
6) Hyperparameters: We used Adam optimizer with learning rate of 0.001.

V. RESULTS

We implemented our method using the Pytorch deep neural network framework. We trained our network with NVDIA-Titanx GPU. Each iteration takes about 4 hours. It takes up to 4 iterations to converge.

A. Performance on Validation Dataset

For every epoch, we trained the model and validated the model on validation dataset (1292 videos). As shown in the graph below, model trained at epoch 3 (model3) performed best on validation set with 99% accuracy. Therefore, we chose model3 for testing.



B. Performance on Testing Dataset

When tested with model3 on testing dataset, the average loss was 0.0243 and accuracy was 99% (6500/6522).

VI. LIMITATIONS AND DISCUSSIONS

The performance on testing dataset is very high. There are two major problems behind this high performance.

1) Three videos from one material class all have same texture. Although there are three videos per one material class, they are videos taken by applying different wind forces to one cloth. They are not videos of different texture cloth of same material. Therefore, there is high possibility of our network learned just the texture, not the change of the wrinkles over time. This leads to the need for visualizing what our network learns at each layer during training session.
2) Dataset for training, validating, and testing were all from same video. Although they are videos made of different frames, but the frames are from same video. To properly test the network, we have to use different videos.

VII. CONCLUSION AND FUTURE WORK

We have presented a learning-baed algorithm to recover material label. Our training videos contain only a single piece of cloth and the recorded cloth is not interacting with any other object. A natural extension would be to learn from videos of cloth directly interacting with the body, under varying lighting conditions and partial occlusions.

REFERENCES

[1] Aubry, Mathieu, Daniel Maturana, Alexei A. Efros, Bryan C. Russell, and Josef Sivic. ”Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3762-3769. 2014.

[2] Battaglia, Peter W., Jessica B. Hamrick, and Joshua B. Tenenbaum. ”Simulation as an engine of physical scene understanding.” Proceedings of the National Academy of Sciences (2013): 201306572.

[3] Bell, Sean, Paul Upchurch, Noah Snavely, and Kavita Bala. ”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.

[4] Bhat, Kiran S., Christopher D. Twigg, Jessica K. Hodgins, Pradeep K. Khosla, Zoran Popovi, and Steven M. Seitz. ”Estimating cloth simulation parameters from video.” In Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation, pp. 37-51. Eurographics Association, 2003.

[5] Borji, Ali, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. ”Salient object detection: A benchmark.” IEEE transactions on image processing 24, no. 12 (2015): 5706-5722.

[6] Bouman, Katherine L., Bei Xiao, Peter Battaglia, and William T. Freeman. ”Estimating the material properties of fabric from video.” In Proceedings of the IEEE international conference on computer vision, pp. 1984-1991. 2013.

[7] Bridson, Robert, Ronald Fedkiw, and John Anderson. ”Robust treatment of collisions, contact and friction for cloth animation.” ACM Transactions on Graphics (ToG) 21, no. 3 (2002): 594-603.

[8] Chen, Wenzheng, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. ”Synthesizing training images for boosting human 3d pose estimation.” In 3D Vision (3DV), 2016 Fourth International Conference on, pp. 479-488. IEEE, 2016.

[9] Cheung, Ernest, Tsan Kwong Wong, Aniket Bera, Xiaogang Wang, and Dinesh Manocha. ”Lcrowdv: Generating labeled videos for simulation-based crowd behavior learning.” In European Conference on Computer Vision, pp. 709-727. Springer, Cham, 2016.

[10] Dai, Jifeng, Kaiming He, and Jian Sun. ”Convolutional feature masking for joint object and stuff segmentation.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3992-4000. 2015.

[11] Davis, Abe, Katherine L. Bouman, Justin G. Chen, Michael Rubinstein, Fredo Durand, and William T. Freeman. ”Visual vibrometry: Estimating material properties from small motion in video.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5335-5343. 2015.

[12] DelPozo, Andrey, and Silvio Savarese. ”Detecting specular surfaces on natural images.” In Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on, pp. 1-8. IEEE, 2007.

[13] Dollr, Piotr, Ron Appel, Serge Belongie, and Pietro Perona. ”Fast feature pyramids for object detection.” IEEE Transactions on Pattern Analysis and Machine Intelligence 36, no. 8 (2014): 1532-1545.

[14] Donahue, Jeffrey, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. ”Long-term recurrent convolutional networks for visual recognition and description.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625-2634. 2015.

[15] Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. ”Rich feature hierarchies for accurate object detection and semantic segmentation.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587. 2014.

[16] Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. ”Region-based convolutional networks for accurate object detection and segmentation.” IEEE transactions on pattern analysis and machine intelligence 38, no. 1 (2016): 142-158.vvvvv

[17] Govindaraju, Naga K., Ilknur Kabul, Ming C. Lin, and Dinesh Manocha. ”Fast continuous collision detection among deformable models using graphics processors.” Computers & Graphics 31, no. 1 (2007): 5-14.

[18] Keskin, Cem, Furkan Kra, Yunus Emre Kara, and Lale Akarun. ”Real time hand pose estimation using depth sensors.” In Consumer depth cameras for computer vision, pp. 119-137. Springer, London, 2013.

[19] Koh, Woojong, Rahul Narain, and James F. O’Brien. ”View-dependent adaptive cloth simulation.” In Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp. 159-166. Eurographics Association, 2014.

[20] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. ”Imagenet classification with deep convolutional neural networks.” In Advances in neural information processing systems, pp. 1097-1105. 2012.

[21] Lee, Huai-Ping, Mark Foskey, Marc Niethammer, Pavel Krajcevski, and Ming C. Lin. ”Simulation-based joint estimation of body deformation and elasticity parameters for medical image analysis.” IEEE transactions on medical imaging 31, no. 11 (2012): 2156.

[22] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. ”Fully convolutional networks for semantic segmentation.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440. 2015.

[23] Miguel, Eder, Derek Bradley, Bernhard Thomaszewski, Bernd Bickel, Wojciech Matusik, Miguel A. Otaduy, and Steve Marschner. ”Datadriven estimation of cloth simulation models.” In Computer Graphics Forum, vol. 31, no. 2pt2, pp. 519-528. Oxford, UK: Blackwell Publishing Ltd, 2012.

[24] Mongus, Domen, Blaz Repnik, Marjan Mernik, and B. alik. ”A hybrid evolutionary algorithm for tuning a cloth-simulation model.” Applied Soft Computing 12, no. 1 (2012): 266-273.

[25] Noh, Hyeonwoo, Seunghoon Hong, and Bohyung Han. ”Learning deconvolution network for semantic segmentation.” In Proceedings of the IEEE international conference on computer vision, pp. 1520-1528. 2015.

[26] Pinheiro, Pedro O., and Ronan Collobert. ”From image-level to pixel-level labeling with convolutional networks.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1713-1721. 2015.

[27] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. ”Faster r-cnn: Towards real-time object detection with region proposal networks.” In Advances in neural information processing systems, pp. 91-99. 2015.

[28] Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang et al. ”Imagenet large scale visual recognition challenge.” International Journal of Computer Vision 115, no. 3 (2015): 211-252.

[29] Shao, Jing, Chen Change Loy, and Xiaogang Wang. ”Scene-independent group profiling in crowd.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2219-2226. 2014.

[30] Shao, Jing, Kai Kang, Chen Change Loy, and Xiaogang Wang. ”Deeply learned attributes for crowded scene understanding.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4657-4666. 2015.

[31] Sharif Razavian, Ali, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. ”CNN features off-the-shelf: an astounding baseline for recognition.” In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 806-813. 2014.

[32] Solmaz, Berkan, Brian E. Moore, and Mubarak Shah. ”Identifying behaviors in crowd scenes using stability analysis for dynamical systems.” IEEE transactions on pattern analysis and machine intelligence 34, no. 10 (2012): 2064-2070.

[33] Syllebranque, Cdric, and Samuel Boivin. ”Estimation of mechanical parameters of deformable solids from videos.” The Visual Computer 24, no. 11 (2008): 963-972.

[34] Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. ”Going deeper with convolutions.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9. 2015.

[35] Wang, Huamin, James F. O’Brien, and Ravi Ramamoorthi. ”Data-driven elastic models for cloth: modeling and measurement.” In ACM Transactions on Graphics (TOG), vol. 30, no. 4, p. 71. ACM, 2011.

[36] Wu, Jiajun, Ilker Yildirim, Joseph J. Lim, Bill Freeman, and Josh Tenenbaum. ”Galileo: Perceiving physical object properties by integrating a physics engine with deep learning.” In Advances in neural information processing systems, pp. 127-135. 2015.