Advances in Artificial Intelligence and Machine Learning; Research 3 (4) 1787-1799 Received 24-10-2023; Accepted 21-12-2023; Published 28-12-2023

Handsight: DeCAF & Improved Fisher Vectors to Classify Clothing Color and Texture With a Finger-Mounted Camera

Alexander J. Medeiros ajmed13@gmail.com
Makeability Lab | HCIL,
University of Maryland
College Park, MD, USA

Lee Stearns lstearns86@gmail.com
Makeability Lab | HCIL,
University of Maryland
College Park, MD, USA

Jon E. Froehlich jonf@cs.umd.edu
Makeability Lab | HCIL,
University of Maryland
College Park, MD, USA

Corresponding Author: Alexander J. Medeiros

Abstract

We demonstrate the use of DeCAF and Improved Fisher Vector image features for clothing texture classification, with a focus on aiding visually impaired individuals in selecting their attire. Choosing clothes based on color and texture is a daily problem for visually impaired individuals. This work is a preliminary attempt to alleviate the problem using a finger-mounted camera and state-of-the-art classification algorithms to identify clothing textures. In evaluating our solution, we used a NanEyeGS camera, small enough to mount on a finger, to collect 520 close-up images across 29 garments, referred to as the HandSight Color-Texture Dataset (HCTD). Secondly, we contribute evaluations of these state-of-the-art recognition algorithms applied to our dataset - achieving an accuracy exceeding 95%. Throughout this article, we review prior research, evaluate our current solution, and outline future project directions.

Keywords: Pattern recognition, Algorithms, Human factors, Visual impairments.

1. INTRODUCTION

Approximately 161 million people globally are visually impaired [1]. To provide visual context to hard-of-sight individuals, our research group has been exploring ways to observe a user's immediate surroundings by augmenting the sense of touch, namely, HandSight. The current work adds to Hand-

Citation: Alexander J. Medeiros, et al. Handsight: DeCAF & Improved Fisher Vectors to Classify Clothing Color and Texture With a Finger-Mounted Camera. Advances in Artificial Intelligence and Machine Learning. 2023;3(4):103.

Sight's capabilities by facilitating the daily task of choosing an outfit. Currently, visually impaired individuals have various adaptations for accomplishing this task including improved organization, high consistency of clothes, or tool-based approaches. Organizational approaches attempt to order clothes where each garment has a dedicated location whereas high consistency avoids the problem by wearing the same style of clothing daily (e.g. black turtleneck and blue jeans). Tool-based approaches include handheld devices and apps such as WayAround which enable individuals to scan stickers or tags affixed to items and hear a user-defined description aloud. However, all these solutions require preparation and foresight while the tool-based approach we introduce, HandSight, allows ad-hoc clothing identification. The primary aim of this research is to build an image processing pipeline that accurately categorizes close-up images of clothing.

We have implemented an image processing pipeline that includes Dense SIFT (DSIFT), DeCAF and Improved Fisher Vectors (IFV) to classify our dataset with >95% accuracy. Moving forward, we will focus on user experience and evaluating different mobile and on-body implementations. We have already worked towards this effort by capturing the HandSight Color-Texture Dataset (HCTD) using a NanEyeGS camera (depicted in FIGURE 1), small enough to mount on the finger (FIGURE 2). The HCTD contains close-up images of clothing under realistic conditions including varied angle, distance, and tautness of fabric.

Figure 1: A NanEye GS camera - only a few millimeters in diameter.

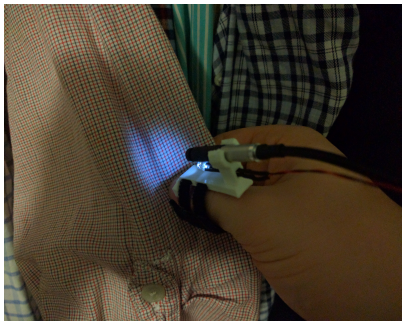


Figure 2: A finger-mounted HandSight prototype with an LED. Notice the shadow casted by the user has little impact on the lit area. By blocking the effects of ambient light, we are able to increase accuracy with consistent lighting.

Recent research in pattern and texture recognition [2], has produced algorithms that reach >98% accuracy on datasets with consistent viewing parameters, while achieving >65% accuracy on datasets such as the DTD [2], and FMD [3]. Our close-up and on-body camera approach generates consistent

image characteristics such that our accuracy remains high even with ambient lighting changes due to the LED fixed to HandSight's finger-mount.Aside from algorithmic related work, Yang et al. has also explored the problem of choosing an outfit with visual impairment [4]. They produced an algorithm for recognizing clothing color and pattern with ~93% accuracy [4]. In contrast, our classification achieves >95% accuracy with a uniquely mobile solution including the finger-mounted camera and LED combo. Lastly, our approach classifies more than twice as many textures as Yang's work.

In summary, our contributions include: (1) the HCTD: a unique set of close-up clothing images; and (2) evaluations of state-of-the-art recognition algorithms applied to our dataset - achieving an accuracy >95%. Our findings further validate the use of DeCAF and IFVs as a means of image classification and provide a novel and difficult image dataset to be used with future HandSight development. Further research is needed to develop a wireless mobile solution that operates in real-time.

2. BACKGROUND

Our texture classification work is largely inspired by the DTD [2], and their success in classifying 47 different textures, including many that accurately categorize clothing patterns such as checkered, striped, floral, etc. Therefore, the 9 categories we chose are derived from the DTD, except for denim and none/solid. The algorithms used for HandSight are adapted from the DTD paper which uses IFVs and DeCAF feature vectors to represent an image. DeCAF vectors have been implemented by the BVLC in the Deep learning framework called Caffe. These were initially used by Krizhevsky et al. (2012) which won the ILSVRC – 2012 [5]. Since that paper, many object recognition tasks have used some variant of the neural network developed in Krizhevsky's work and found satisfactory results [5]. Furthermore, the DTD evaluated several different feature vector collections including IFV, BOVW, VLAD, LLC, KCB, and DeCAF. They found their highest accuracy with IFV + DeCAF on every dataset they experimented with. Therefore, we went directly to IFV + DeCAF.

In terms of the physical design of HandSight, we used one of our group's prototypes which is evaluated in [6]. In short, HandSight is a finger mountable set of sensors allowing a user to gather visual context through the intuitive sense of touch. We have already studied HandSight's use for reading text on a page and using gestures as input control for a mobile device as depicted in FIGURE 3 [6, 7]. The current work further explores the ability of finger-mounted sensors to facilitate daily tasks such as choosing clothes.

Finally, most similar to the current work, Yang et al. has implemented a system that can classify 4 patterns and 11 clothing colors with ~93% accuracy [4]. They implemented a speech-to-text controller for sending commands to the system such as "start recognition" and "turn off system." Moreover, they completed a user-study which concluded that blind users desired such a system to support more daily independence [4]. In comparison, the benefits of our approach are evident in low-lighting conditions and in general use. If there is low-lighting for Yang's system, the recognition has limited capability, however, since we've outfitted the camera with an LED, our approach is generally not subject to ambient light. Secondly, Yang's system requires the user to position the clothing to occupy the camera's full view. To our knowledge, HandSight is the first to use close-up and local features for general classification of clothing color and texture.

Figure 3: One of many HandSight prototypes which facilitates reading for the visually impaired.

Figure 4: Many color grabbers only compute average color rather than the few that are most prevalent in an image.

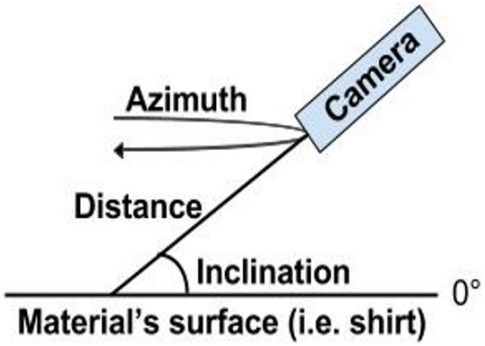


Figure 5: Definition of three columns of the HCTD: distance, inclination, and azimuth.

3. METHODS: HANDSIGHT COLOR-TEXTURE DATASET

To evaluate our approach, we needed a dataset representative of our problem domain (close-up, LED lit, and lower resolution camera). We call this the HandSight Color-Texture Dataset (HCTD). For each garment, we strapped the fabric to a board with binder clips or placed it on a hanger to vary the tautness. Then we placed the camera in position using a combination of rulers and protractors before capturing the view of the NanEyeGS camera. Upon capture we would enter the parameters

into the csv and set the image's filename based on the new row's image ID. To maintain realistic conditions, there is no post-processing of the images. The dataset is tabulated as a csv file with the following 11 columns for each image:

1. Image ID
2. Label
3. Label ID
4. Distance between camera and material
5. Angle of camera to material (inclination)
6. Orientation of camera (azimuth)
7. Scale (pixels per cm (PPCM))
8. Lighting (LED configuration)
9. Tension of material (e.g. taut or hanging)
10. Notes
11. Colors

The image ID (1) is an integer primary key and (2,3) are the label and label ID (e.g. checkered and 0). Distance, inclination, and azimuth (4,5,6) are demonstrated in FIGURE 5, and shown in practice in FIGURE 6. We included these variations to keep our training set angle, distance, and tension agnostic. Although the current work only focuses on close-up images, distance is included because we intend to test a wrist-mounted solution in the future. The scale (7) is the width of the camera's resolution (640 pixels) divided by the number of centimeters measured across the viewed object's horizontal. Scale is included because it allows others to use different camera configurations and maintain the same view scale. As depicted in FIGURE 8, the camera can pick out individual threads of the clothing. That level of detail fades in images taken at 12 cm away (wrist mounted distance). The lighting (8) is a measure of the power supplied to the LED (0-255 times the 5V power source squared divided by 10 ohms). This measure will prove useful in future work when we automate the LED brightness. The tension (9) of the material is important because of the variability seen in normal conditions such as hanging in a closet or laying in a drawer. Finally, the notes (10) are annotations for each image and the colors (11) are manually labeled colors of each garment.

As depicted in FIGURE 6, the variables and parameters chosen result in 16 possible configurations for each article of clothing considering 2 azimuth angles, 2 distances, 2 inclination angles, and 2 tensions. Notice the difference between the detail of fabric in FIGURE 7, and FIGURE 8. This is one example of why the HCTD is a novel dataset with respect to clothing texture. For reference, the HCTD includes an image like FIGURE 7, for each article of clothing.

In FIGURE 8, we have 9 texture categories which were derived from previous work in categorizing texture [2, 8]. Using the previous work and eliminating all texture words such as bubbly, honeycombed, etc. (textures that do not describe clothing). We enumerated our texture categories. We also combined texture categories such as striped, banded, and lined into a single category called striped. Further, we added denim and none to compensate for common clothing categories that did not exist in previous texture categorization work.

The dataset has 520 images across the 9 texture types and 29 distinct articles of clothing, as outlined in TABLE 1.

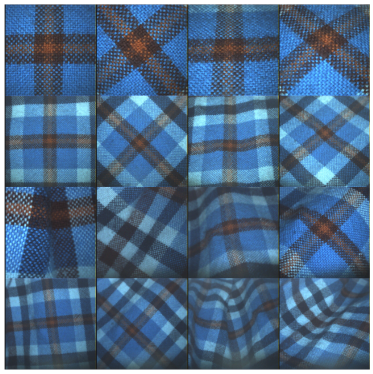


Figure 6: The 16 possible configurations considering 2 azimuth angles, 2 distances, 2 inclination angles, and 2 tensions. The 1st and 3rd row are at a distance of 5cm while the others are at 12cm. The 2nd and 4th columns are at an azimuth angle of 45°, while the others are at 90°. The 1st and 2nd rows are taut; the others are hanging on a hook. Finally, the 3rd and 4th columns are at an inclination of 45°; the others are at 90°.

Figure 7: This is the original texture of FIGURE 6, captured with a Nexus 5X phone camera. The dataset includes an image similar to this for each texture used.

The HCTD is similar in function to the DTD [2], and the CCNY Clothing Dataset [2, 4], however the HCTD is novel in its close-up images of the most common clothing texture types under varied and realistic conditions. Our dataset should prove useful for any research involving clothing texture recognition. Especially considering most local clothing features extend to the entire piece of clothing. In other words, most clothes can be identified by only a few centimeters of the article.

4. METHODS: TEXTURE CLASSIFICATION PIPELINE

In this section, we present our methodology for clothing texture classification, inspired by insights from the Describable Textures Dataset (DTD) [2]. Our objective was to achieve precise texture classification for clothing using three key image feature extraction techniques: DeCAF, Dense SIFT

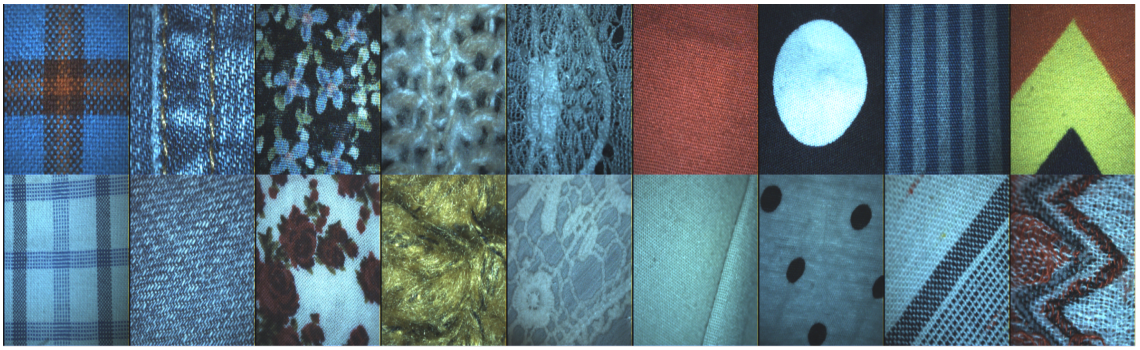


Figure 8: Depicts each texture in the HCTD. Notice the significant variability between samples of the same texture. Also note the similarities between textures such as denim and none or the second row of none and striped.

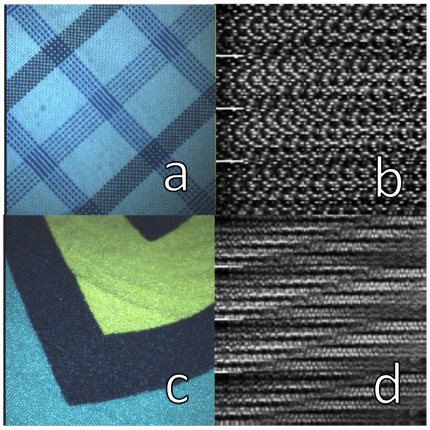


Figure 9: (a,c) are checkered_00038 and zigzagged_00367 images from the HCTD. (b,d) are the first of ten output matrices of the DenseSIFT algorithm, illustrated as an intensity map to demonstrate the differences between the two images.

(DSIFT), and Improved Fisher Vectors (IFV). We provide an overview of these techniques in the context of our study.

The first step of our image feature extraction process is obtaining DeCAF descriptors through a Deep Convolutional Neural Network[1]. Traditional neural network training demands extensive computational resources and large datasets. To overcome these limitations, we use a technique called feature extraction using a pre-trained AlexNet Caffe model. This approach allows us to obtain feature vectors from our garment images without the need for resource-intensive training. FIGURES 10 and 11, provide visual representations of the DeCAF descriptor extraction process. These plots offer insights into the information encapsulated within each DeCAF descriptor. It is important to note that while AlexNet was originally trained on unrelated images, its output vectors remain useful for our specific application. The early layers of the neural network capture general

[1] We used the Caffe Deep Learning Framework.

Table 1: Outlines basic stats of the HCTD. Under the first column is each texture's label ID, followed by the texture type, number of images per texture, and the number of garments for that texture.

LabelID	Texture type	# images	# articles
0	Checkered	88	5
1	Denim	40	3
2	Floral	88	4
3	Knitted	32	2
4	Lacelike	48	2
5	None	48	3
6	Polka-dotted	48	3
7	Striped	64	4
8	Zigzagged	64	3
	TOTAL	**520**	**29**

patterns like edges and blobs, while the later layers specialize in finer details. By omitting the bottom layer, we obtain compact 4096-dimensional DeCAF vectors that capture essential image characteristics.

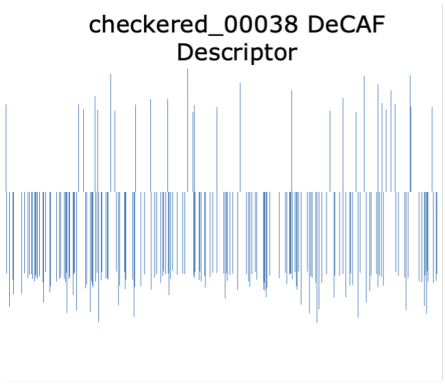


Figure 10: The 4096-D DeCAF feature vector which is output from the pre-trained AlexNet Caffe model after employing feature extraction on the HCTD image, checkered_00038.

Secondly, our texture classification pipeline incorporates the Dense Scale Invariant Feature Transform (DSIFT) algorithm[2], used for its ability to capture image features resilient to scale changes and its focus on blobs and corners. FIGURE 9 (b,d) illustrates the DSIFT process, showcasing intensity maps where each column corresponds to a 128-dimensional SIFT descriptor. DSIFT achieves scale invariance by iteratively resizing images and extracting descriptors, a process controlled by specific parameters[3]. FIGURE 9 visually depicts the signals captured by DSIFT. Our parameterization results in matrices measuring 128 by 392,584 elements while future research may explore parameter selection to optimize performance while preserving classification accuracy [5].

[2] We used the VLFEAT implementation of fast Dense SIFT.
[3] See source code for complete implementation and choice of parameters.

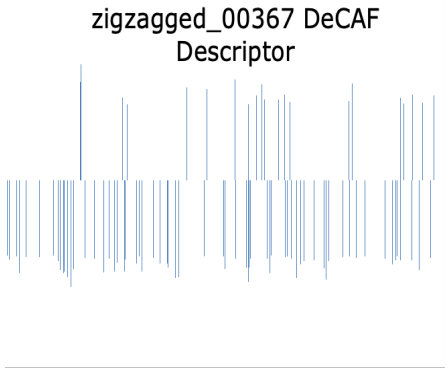


Figure 11: 4,096-D DeCAF vector after feature extraction of HCTD image, zigzagged_00367.

The final component of our texture classification pipeline is the Improved Fisher Vector (IFV), obtained by pooling DSIFT matrices and depicted in FIGURES 12, and 13. The IFV algorithm refines the DSIFT matrix into a 40,960-dimensional vector. When combined with the 4,096-dimensional DeCAF vector, it serves as the input for our Support Vector Machine (SVM) classifier. Powered by the Dual Stochastic Coordinate Ascent (SDCA) strategy, this classifier leverages the composite feature vector to discern patterns within clothing textures.In summary, our texture classification pipeline relies on three components - DeCAF, DSIFT, and IFV - to enable categorization of clothing textures. Each component contributes unique features, consistently achieving accuracy levels exceeding 95%. This methodology forms the foundation of our research and contributes to advancements in clothing texture classification.

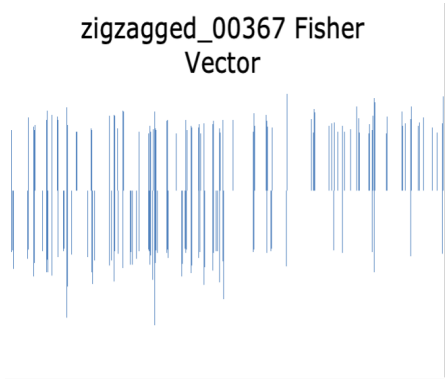


Figure 12: 40,960-D IFV after pooling the DSIFT features.

5. RESULTS

Our classification approach involved training an SDCA SVM on the normalized IFV+DeCAF feature vectors and evaluating accuracy across 40-fold random subsampling. This process varied the percentage of HCTD used for training, ranging from 20% to 80%. The results are presented in

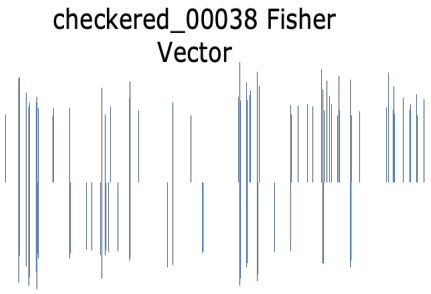
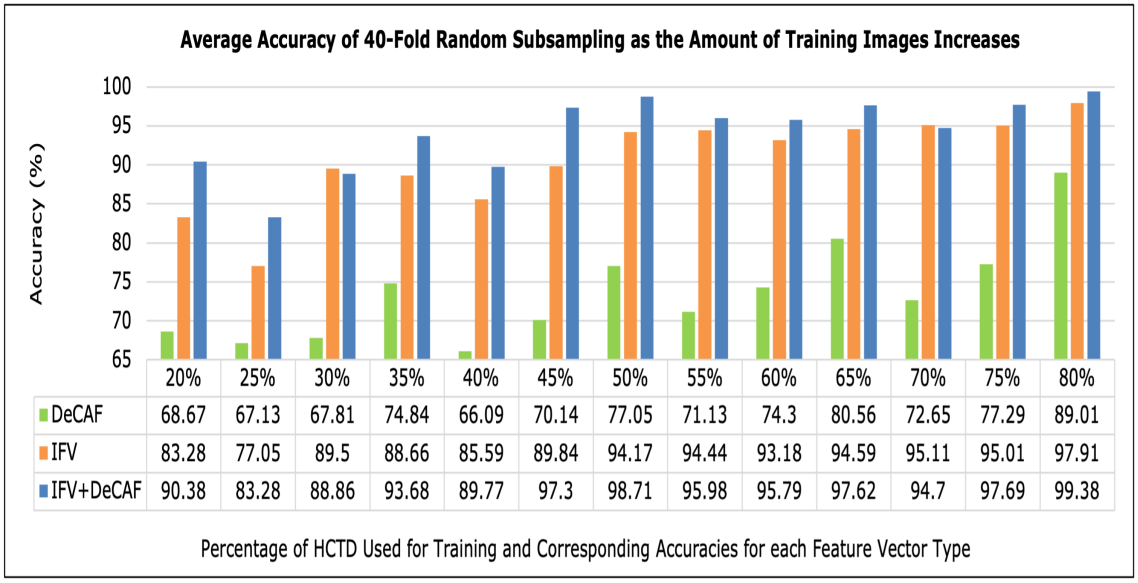


Figure 13: 40,960-D IFV after pooling the DSIFT features.

FIGURE 14, with three sets of bar graphs representing DeCAF-only vectors, IFV-only vectors, and the combined IFV+DeCAF approach.



	20%	25%	30%	35%	40%	45%	50%	55%	60%	65%	70%	75%	80%
DeCAF	68.67	67.13	67.81	74.84	66.09	70.14	77.05	71.13	74.3	80.56	72.65	77.29	89.01
IFV	83.28	77.05	89.5	88.66	85.59	89.84	94.17	94.44	93.18	94.59	95.11	95.01	97.91
IFV+DeCAF	90.38	83.28	88.86	93.68	89.77	97.3	98.71	95.98	95.79	97.62	94.7	97.69	99.38

Figure 14: Depicts experimental results when using our classifier on the HCTD. There are three sets of bars here: Green, orange, and blue. Each bar represents the average accuracy of our classifier across 40 trials at each 5% interval of the HCTD used as the training set. Green represents using only the DeCAF feature vectors as input for the SVM. Orange bars represent the IFV, and blue represents IFV+DeCAF.

Notably, DeCAF-only classifications exhibited relatively low accuracy (green bars), contrary to expectations set by the DTD's >90% accuracy using DeCAF on other datasets [2]. A closer examination revealed the HCTD, with only eight images per garment and single-digit garments per class, lacked sufficient data to accurately classify texture using DeCAF's smaller 4,096-dimensional

vectors. A higher volume of training data is necessary to compensate for DeCAF's relative lack of information per image.

Analyzing FIGURE 14, several observations emerge. Each 5% increment corresponds to 2-4 images per texture class, and a general upward trend is noticeable: as the percentage of HCTD used for training increases, accuracy improves. There are dips at 25%, 40%, 55%, and 70%, seemingly reflecting a poorly behaved function. The dip at 25% stands as an outlier, with the 30% accuracy figure merely 0.4% away from outlier status. The IFV-only trials (orange bars) exhibit unstable accuracy until approximately 40% dataset usage. Beyond this point, IFV accuracy stabilizes and follows a predictable upward trajectory. Conversely, DeCAF-only remains unpredictable throughout all 13 increments of HCTD usage, emphasizing the need for more data to use the smaller DeCAF vectors effectively. As expected, the composite vectors (IFV+DECAF) performed the best with a stable >95% accuracy for all trials past 40% dataset usage. These are encouraging results, indicating the pipeline can classify clothing texture with less than a majority of the dataset used for training.However, in practical scenarios, such as those involving visually impaired individuals using our HandSight system, a small margin of error can have meaningful consequences. A classification error could lead to an incorrect choice of clothing, potentially causing discomfort or embarrassment. Therefore, while our system's overall accuracy is promising, there is room for refinement to reduce the error rate further.

These findings underscore the critical role of data volume in achieving consistent and accurate classifications. While our system's performance is promising, it also highlights the need for more extensive data collection efforts to provide a richer training dataset. By acquiring a more substantial number of garments and variations in perspective, we can aim for human-level accuracy across diverse sets of clothing, bringing us closer to the system's practical viability.

6. DISCUSSION AND CONCLUSION

Our work in clothing texture classification, inspired by the challenges faced by approximately 161 million visually impaired individuals globally [1], represents progress toward augmenting the sense of touch and facilitating everyday tasks, such as choosing an outfit, through our HandSight system. In contrast to existing methods that require meticulous organization or clothing consistency, HandSight offers a unique ad-hoc approach to clothing identification.

In comparison to current standards and alternative methods employed to aid visually impaired individuals in clothing selection, our approach distinguishes itself by its potential real-time adaptability and high accuracy. While many existing solutions necessitate tedious preparation and foresight, HandSight empowers users with on-the-fly clothing identification. This distinction is particularly relevant for individuals who seek a seamless and spontaneous experience in selecting their attire.Our image processing pipeline, employing techniques inspired by the DTD, consistently achieves accuracy levels exceeding 95%. However, a notable challenge lies in achieving real-time performance, as our current pipeline operates on the order of seconds per image with an Intel i7 processor and 8GB of RAM. While our system excels in accuracy, enhancing its speed remains a critical avenue for future development.

Despite our achievements, several limitations persist. The ”close-up image drawback,” exemplified in FIGURE 15, underscores the importance of global clothing color and texture classification in contrast to the current solution. For example, describing a green, pink, and white floral skirt based solely on local color and texture information falls short of conveying its full visual essence to a visually impaired user.

Figure 15: An example garment that requires more elaborate feedback than mere color and texture for communicating to a user.

The HandSight Color-Texture Dataset (HCTD), while a valuable resource, currently remains relatively small. To build a more comprehensive classifier, we recognize the need for a more extensive dataset, comprising hundreds of garments and encompassing a wider array of perspective variations. Future work involves implementing an automated data collection solution, incorporating a robotic arm for precise camera orientation adjustments. Additionally, collaboration with local thrift stores to expand our clothing variety could contribute significantly to dataset enrichment. To avoid more elaborate technical solutions, we could also explore crowdsourcing close-up garment images to provide a sufficiently diverse clothing set and reduce demographic or locale biases.

In terms of potential improvements and future research directions, fine-tuning the Caffe model with the HCTD dataset holds promise for enhancing accuracy without compromising performance. Transitioning to Caffe as the primary framework is expected to become feasible with a larger dataset, potentially enabling real-time performance. Currently, DeCAF processing takes approximately 180 milliseconds per image, while IFV demands approximately 3.8 seconds per image.In summary, our work addresses the critical challenge of clothing texture recognition for visually impaired individuals. We have presented a dataset and classification approach that consistently achieves accuracy levels exceeding 95%, offering a significant step towards providing visual context to those with visual impairments. While challenges and limitations remain, our research lays a strong foundation for future advancements, ultimately bridging the gap between computer vision and enhancing the daily lives of individuals with visual impairments.

As we continue to develop HandSight, we contribute to a more inclusive and accessible world, one where visually impaired individuals can participate in spontaneous and independent clothing selection.

References

[1] Resnikoff S, Pascolini D, Etya'ale D, Kocur I, Pararajasegaram R, et al. Global Data on Visual Impairment in the Year 2002. Bull World Health Organ. 2004;82:844-851.

[2] Cimpoi M, Maji S, Kokkinos I, Mohamed S, Vedaldi A, et al. Describing Textures in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014:3606-3613.

[3] Sharan L, Rosenholtz R, Adelson EH. Material Perception: What Can You See in a Brief Glance? J Vis. 2009;09:784.

[4] Yang X, Yuan S, Tian YL. Assistive Clothing Pattern Recognition for Visually Impaired People. IEEE Trans Hum Mach Syst. 2014;44:234-243.

[5] Krizhevsky A, Sutskever I, Hinton GE. Imagenet Classification With Deep Convolutional Neural Networks. Advances in neural information processing systems. 2012;25.

[6] Stearns L, Du R, Oh U, Jou C, Findlater L, et al. Evaluating Haptic and Auditory Directional Guidance to Assist Blind People in Reading Printed Text Using Finger-Mounted Cameras. ACM Trans Access Comput. 2016;9:1-38.

[7] Stearns L, Oh U, Cheng BJ, Findlater L, Ross DA, et al. Localization of Skin Features on the Hand and Wrist From Small Image Patches. In: Proceedings of the ICPR. 2016:1003-1010.

[8] Bhushan N, Rao AR, Lohse GL. The Texture Lexicon: Understanding the Categorization of Visual Texture Terms and Their Relationship to Texture Images. Cogn Sci. 1997;21:219-246.