FashionFinder: Efficient Clothing Retrieval with Deep Features and Vocabulary Trees

Om Mali, Naga Kushal Ageeru
Khoury College of Computer Sciences
Northeastern University
Boston, MA 02120

1 Introduction

Images are a powerful form of data with extensive applications across fields such as object recognition, medical imaging, and recommendation systems. However, they come with significant storage and computational challenges, particularly when fast and accurate recommendations are required. Content-Based Image Retrieval (CBIR) systems address this by analyzing visual features like shape, color, and texture to identify similar images efficiently. One prominent method in CBIR is the vocabulary tree approach, which combines hierarchical clustering and the bag-of-features model. Vocabulary trees organize image features into nodes that represent high-level characteristics, enabling fast retrieval and robust performance. Each node in the tree contributes to image classification and comparison, allowing efficient processing of large datasets.

This project explores implementing vocabulary trees to recommend similar images within a clothing dataset. It involves training on labeled images, testing with internal and external inputs, and optimizing retrieval through various feature descriptors. Additionally, the project evaluates the impact of different descriptors on MAP and other evaluation metrics, offering insights into the strengths and tradeoffs of this approach in CBIR systems.

Code citations:

1. Rustworkx: Rustworkx (n.d.). *Rustworkx: A Python library for working with graphs*. Retrieved from `https://www.rustworkx.org/`
2. OpenCV ORB Detector: OpenCV (n.d.). *cv2.orb – ORB Feature Detection*. Retrieved from `https://docs.opencv.org/4.x/db/d95/classcv_1_1ORB.html`
3. OpenCV SIFT Detector: OpenCV (n.d.). *cv2.sift – SIFT Feature Detection*. Retrieved from `https://docs.opencv.org/4.x/d7/d60/classcv_1_1SIFT.html`
4. ResNet (PyTorch): PyTorch (2015). *ResNet-50 model*. Retrieved from `https://pytorch.org/hub/pytorch_vision_resnet/`
5. SuperPoint: Magic Leap Community (2020). *SuperPoint: A deep learning model for keypoint detection and description*. Hugging Face. Retrieved from `https://huggingface.co/magic-leap-community/superpoint`

2 Literature Survey

The project draws upon insights from several pivotal papers in the field of computer vision.

Nister and Stewenius [2006] introduces an efficient method for large-scale object recognition using hierarchical vocabulary trees. Their approach employs hierarchical k-means clustering on image features, enabling rapid and accurate image retrieval—a fundamental component of this project.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

D. Lowe[2004] introduces the Scale-Invariant Feature Transform (SIFT). SIFT detects robust local image features invariant to scaling, rotation, and partial illumination changes, enabling reliable image matching across varied conditions. It laid the foundation for feature-based computer vision applications such as object recognition and image stitching.

Margarita Bratkova et al. [2009] A Practical Opponent Color Space for Computer Graphics explores the development of the oRGB color space, a model inspired by the human visual system's opponent processing. Unlike traditional models like RGB, oRGB separates chromatic and achromatic components, providing advantages in applications like color manipulation, rendering, and color blending. The authors focus on its practical implementation, ensuring it is computationally efficient and compatible with existing graphic systems, making it a viable choice for enhancing color processing in computer graphics

Ethan Rublee et al.[2012] proposes a novel feature detection and description method combining the FAST keypoint detector with the BRIEF descriptor, making it computationally faster than both SIFT and SURF. ORB is designed for real-time applications requiring efficiency while maintaining robustness against rotation and scale variations. By leveraging a multi-scale image pyramid, it overcomes the computational expense of SIFT and SURF, making it ideal for resource-constrained devices. The paper shows that ORB can provide competitive results in terms of feature matching and object recognition tasks while being highly efficient.

Bo Cheng, Li Zhuo, Pei Zhang, and Jing Zhang [2014] explores an efficient method for large-scale image retrieval using vocabulary trees. The authors propose a hierarchical approach to organize image features into a tree structure, enabling faster image retrieval by leveraging compact representations of visual data. This method improves scalability and search efficiency compared to traditional techniques, making it highly suitable for large image databases in applications such as content-based image retrieval (CBIR). The paper emphasizes the balance between accuracy and speed, offering a practical solution for real-time image recognition tasks.

Kaiming He et al.[2015] introduces deep residual learning, enabling neural networks to train ultra-deep architectures by mitigating the vanishing gradient problem through shortcut connections. This innovation significantly advanced state-of-the-art performance in image recognition tasks like ImageNet classification and object detection.

Olivier Jeunen et al[2020] explores the effectiveness of Normalized Discounted Cumulative Gain (nDCG) for evaluating recommendation systems. It examines the metric's robustness in off-policy contexts and its ability to capture user satisfaction by prioritizing high-relevance items in top-ranked positions. The study highlights the suitability of nDCG for assessing ranking quality in recommendation tasks, addressing challenges like bias and relevance weighting, and providing insights into improving system performance

3 Methodology

Dataset

The dataset used for this project consists of approximately 1,100 high-resolution images of garments, covering various types, colors, and designs. Of these, 1,000 images were selected for training and experimentation with the model. The dataset is organized into distinct categories, which are detailed below. The methodology is divided into two phases: the offline phase, where model training and preparation occur, and the online phase, which focuses on real-time model testing and retrieval. The dataset can be found at
`https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-dataset`

Methodology

Offline Phase: In the offline phase, local features are extracted from the dataset of images and incorporated into a vocabulary tree. During the tree's formation, these feature descriptors are clustered into visual words using clustering algorithms, creating the vocabulary structure. The tree structure is constructed by recursively partitioning the feature space into respective clusters and

Table 1: Garment Type Distribution

Garment Type	Training	Testing
T-Shirts	125	20
Jeans	125	20
Sweaters	125	20
Shirts	125	20
Dresses	125	20
Shorts	125	20
Casual Shoes	125	20
Skirts	125	3

Table 2: Color Distribution

Color	Training	Testing
Blue	304	45
Black	192	30
White	106	20
Grey	83	12
Red	66	9
Purple	55	8
Green	55	6
Brown	51	5
Pink	50	3
Beige	14	2
Yellow	12	2
Orange	7	1
Multi	5	1

repeating the process for each cluster, thereby capturing image information at various levels of abstraction. Once the tree is created, an inverted index is generated at each leaf node, which records the number of times each image traversed that node during its insertion. The inverted indexes for internal nodes are not stored at the nodes and are calculated by performing an union operation over inverted indexes of its children, resulting in efficient usage of space. Using this inverted index, a sparse 2D vector of size (number of images) × (number of nodes in the tree) is generated. This vector describes the frequency with which each image traveled through specific nodes, where each node is weighted by the number of images that passed through it.

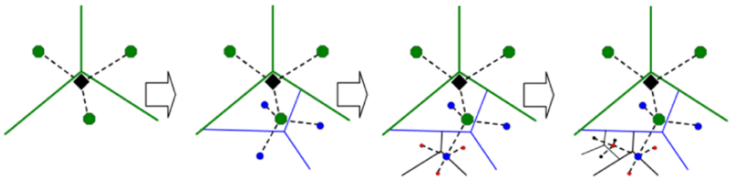

Figure 1: An illustration of the process of building the vocabulary tree.

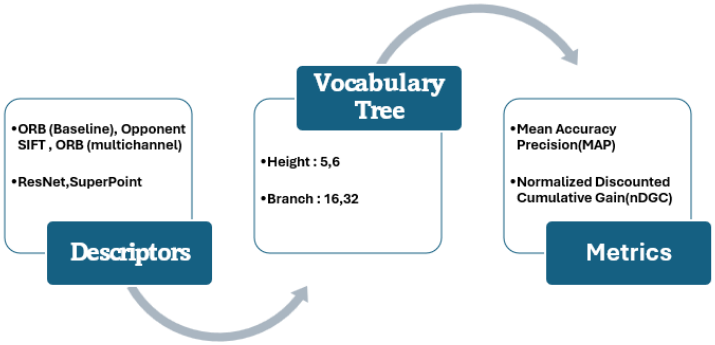

Figure 2: Flowchart and Combinations at Each Phase

Online Phase: During the online phase, an incoming query image is processed by detecting key points and extracting features. Each feature's path through the vocabulary tree is recorded, resulting in a sparse bag-of-words vector of size equal to the number of nodes in the tree for the query image. This vector is weighted by the node weights assigned during tree creation and compared with the bag-of-words vectors of the training images using Minkowski distance to generate similarity scores

for the corresponding images in the training dataset. The process ensures efficient retrieval, as it requires only a 1D vector comparison to match images. Furthermore, the distance is calculated only for nodes that are non-zero in both the query and training image vectors, optimizing computational performance. By traversing the tree, the system efficiently identifies the most similar images in the database, providing fast retrieval based on feature similarity.

Descriptors

ORB (Baseline) : ORB is a fast, lightweight descriptor that combines the FAST detector with the BRIEF binary descriptor, focusing on speed and rotation invariance for real-time applications. While it is computationally efficient, its accuracy in feature matching is lower compared to SIFT, particularly in highly textured or blurred images, and it lacks robustness to significant scale changes.

Opponent SIFT : Opponent SIFT adapts SIFT to color imagery by working on opponent color channels, blending spatial and color information for more robust feature representation in color-sensitive datasets. This enhancement improves performance in scenarios where multi-color objects dominate. However, the method introduces added computational complexity and can struggle with extreme variations in lighting or requires precise parameter tuning for different color spaces.

ORB (multi channel): ORB (multi-channel) is an adaptation of the ORB feature detection and description algorithm designed to work across multiple color channels, enhancing its robustness for colored images. By processing features from each channel separately and combining them, this approach ensures improved performance in scenarios where color plays a critical role, such as garment image retrieval.

Superpoint : SuperPoint is a deep learning-based feature detector and descriptor that combines keypoint detection and feature description into a single end-to-end framework. It uses a fully-convolutional network to identify interest points and their corresponding descriptors in images. In the context of a vocabulary tree for CBIR, SuperPoint's high-quality, repeatable keypoints improve the effectiveness of image retrieval systems by providing strong, consistent feature representations.

Resnet : ResNet introduces skip connections to bypass layers, solving the vanishing gradient problem and enabling deeper networks for enhanced feature extraction. For garments, its capacity to extract hierarchical features like edges, textures, and finer details ensures high accuracy in identifying subtle similarities between clothing items. ResNet performs particularly well when trained on datasets resembling its garment-based target domain, leveraging learned textures and patterns for robust garment feature representation.

Methodology for Feature Extraction Using ResNet

The methodology involving ResNet begins by passing the training images through the ORB algorithm, which detects keypoints in each image. These keypoints are then used as input to the neural networks (ResNet). The pre-trained weights of these networks are applied to the keypoints, allowing them to generate robust feature descriptors. These descriptors are further used for retrieval, by leveraging the learned hierarchical features and representations from the network.

Metrics and Evaluations

Mean Average Precision (MAP) : It is a metric commonly used to evaluate the performance of Content-Based Image Retrieval (CBIR) systems. It calculates the average precision of retrieval results across multiple queries, considering the ranking of relevant images. MAP emphasizes the importance of retrieving relevant images at higher ranks, offering a more comprehensive evaluation of a system's effectiveness . When we refer to MAP color, it means that the ground truth values used for classification are derived from the color columns. This terminology also extends to other columns and metrics.

Normalized Discounted Cumulative Gain (nDCG) : It accounts for the position of relevant images in the result list by giving higher weights to relevant images ranked higher. nDCG normalizes the cumulative gain to ensure comparability across queries, emphasizing ranking quality over just retrieval accuracy.

4 Experiments

We have feature descriptors of various types and vocabulary trees with depths of 5 and 6, and branch factors of 16 and 32. Our next steps are to fit the model with these configurations, evaluate test performance, and compare results. The experiments aim to assess the retrieval accuracy of feature extractors (SIFT, ORB, SuperPoint, ResNet) and analyze the impact of tree depth and branch factor, identifying the most effective method for garment image retrieval.

Based on the metrics, following conclusions can be concurred:

- The SuperPoint keypoint descriptor outperforms other methods across all metrics, particularly in "Article Type MAP" and "Article Type NDCG". This showcases its higher ability for detecting the texture and shape of the image.
- The ORB + ResNet and Multi Channel ORB models perform well but lag behind SuperPoint. They show balanced performance across article type, color type, and combined labels.
- SIFT with Opponent Color Spaces performs the worst among all models incuding the baseline. It also takes the most amount of time to fit the dataset in vocabulary tree.
- Branches and Depth have weak correlations with MAP and nDCG metrics, indicating that increasing these parameters does not necessarily improve retrieval quality significantly. Optimal values may lie in a balanced range, avoiding overly deep or wide trees.
- Color and Article Type Metrics are consistently lower than individual label metrics, highlighting the challenge of multi-label retrieval.
- Color Only metrics are lower than Garment Only metrics for all models indicating the shortcomings of these models while dealing with color similarities.
- Neural Network models were trained using a GPU. If a GPU is not available, then the Multi Channel ORB model with 32 branches and max depth of 6 can be used.

5 Conclusion

The vocabulary tree method for retrieving clothing images is an effective way to organize and search through large datasets, but its performance depends a lot on the type of feature descriptors and tree settings used. Advanced descriptors like SuperPoint performed much better than older methods like SIFT and ORB because they can capture more detailed and meaningful features that match the complexity of clothing images. Descriptors like ORB + ResNet offered a good balance by being fast to compute while still delivering strong results, making them a good option when resources are limited.

Looking at the structure of the vocabulary tree, including its branching factor and depth, we found that these had little impact on accuracy and ranking beyond a certain point. While these settings are important for determining how fast and scalable the tree is, they didn't significantly affect the system's retrieval performance. The model could be improved by experimenting with smaller values for branching factors and depths to identify optimal configurations.

A noticeable issue arose in tasks that used more than one label, like combining color and article type. The system didn't perform as well in these cases, showing it struggles to handle complex, multi-label information. This means that while the method works well for simpler tasks using just one label, it needs further improvements to manage more complicated scenarios effectively.

Given more time, we would like to explore different keypoint detectors and feature descriptors, and experiment with reducing background noise present in images. We are also interested in developing an image segmentation model that segments clothes from an image and uses the vocabulary tree to retrieve cloth images similar to segmented clothes images.

Table 3: Evaluation Metrics for different Models. (B: Branch Factor, H: Max Depth of Tree)

Model	**Garment Only MAP**	**Garment Only nDCG**	**Color Only MAP**	**Color Only nDCG**	**Garment and Color MAP**	**Garment and Color nDCG**
Baseline ORB B:16, H:5	0.660283	0.588763	0.470736	0.401739	0.262372	0.218716
SIFT Opponent Color Spaces B:32, H:6	0.600697	0.537889	0.479083	0.407362	0.272080	0.233683
Multi Channel ORB B:16, H:5	0.722999	0.654031	0.525283	0.445299	0.316590	0.272290
Multi Channel ORB B:32, H:6	0.745825	0.668289	0.520955	0.447883	0.345433	0.294775
ORB + Resnet B:16, H:5	0.781414	0.732964	0.445472	0.385965	0.336073	0.298232
ORB + Resnet B:32, H:6	0.766284	0.720678	0.451043	0.386878	0.327223	0.280701
SuperPoint B:16, H:5	0.881503	0.840103	0.516941	0.445794	0.445833	0.379740
SuperPoint B:16, H:6	0.885290	0.849106	0.492790	0.418230	0.433488	0.373203
SuperPoint B:32, H:5	0.863847	0.826496	0.499295	0.424136	0.440780	0.377700
SuperPoint B:32, H:6	0.887304	0.856362	0.502692	0.423873	0.425910	0.366735

References

Nister, D., Stewenius, H. (2006). Scalable recognition with a vocabulary tree. Computer Vision and Pattern Recognition (CVPR), 2006 IEEE Computer Society Conference, 1–8. https://doi.org/10.1109/CVPR.2006.75

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

Boulos, S., Pellacini, F., Meyer, G. W. (2009). oRGB: A practical opponent color space for computer graphics. IEEE Computer Graphics and Applications, 29(1), 42–55. https://doi.org/10.1109/MCG.2009.3

Rublee, E., Rabaud, V., Konolige, K., Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. International Conference on Computer Vision (ICCV), 2564–2571. https://doi.org/10.1109/ICCV.2011.6126542

Cheng, B., Zhuo, L., Zhang, P., Zhang, J. (2014). Large-scale image retrieval based on the vocabulary tree. Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2014.469

He, K., Zhang, X., Ren, S., Sun, J. (2015). Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90

Vasileva, D., Golodetz, S., Hogg, D. (2018). Fashion Compatibility through Type-Aware Embeddings. Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). https://doi.org/10.1109/WACV.2018.00078