Closing the Gap Between Query and Database through Query Feature Transformation in C2C e-Commerce Visual Search

Takuma Yamaguchi, Kosuke Arase, Riku Togashi, Shunya Ueta
Mercari, Inc
Tokyo, Japan
{kumon,kosuke.arase,riktor,hurutoriya}@mercari.com

ABSTRACT

This paper introduces an image representation technique for visual search on a consumer-to-consumer (C2C) e-commerce website. Visual searching at such websites cannot effectively close the gap between query images taken by users and database images. The proposed technique consists of extraction of a lightweight deep CNN-based feature vector and transformation of a query feature. Our quantitative and qualitative experiments using datasets from an online C2C marketplace with over one billion items show that this image representation technique with our query image feature transformation can improve users' visual search experience, particularly when searching for apparel items, without negative side effects on nonapparel items.

CCS CONCEPTS

• **Information systems** → **Image search**;

KEYWORDS

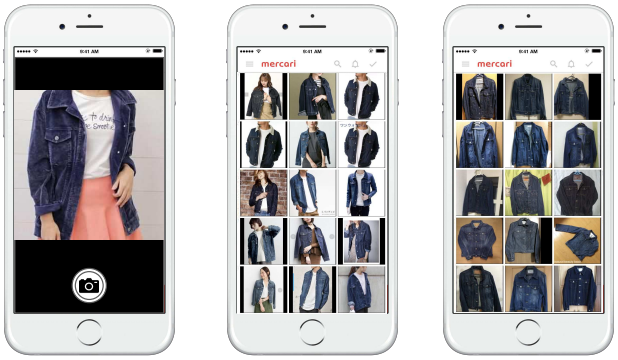
Content-based Image Retrieval, Deep Learning, e-Commerce

ACM Reference Format:
Takuma Yamaguchi, Kosuke Arase, Riku Togashi, Shunya Ueta. 2019. Closing the Gap Between Query and Database through Query Feature Transformation in C2C e-Commerce Visual Search. In *Proceedings of the SIGIR 2019 Workshop on eCommerce (SIGIR 2019 eCom)*, 4 pages.

1 INTRODUCTION

The explosive increase of online photos, driven by social networking and e-commerce sites, has focused researchers' attention on visual search, also called content-based image retrieval [9–11]. Many newly posted photos are listed on consumer-to-consumer (C2C) e-commerce sites, where most sellers are not professional photographers or retailers; therefore, buyers are often stymied by the poor quality or limited quantity of item information and keywords. Moreover, buyers might not even know the correct keywords to use to find their desired items. In such a situation, image-based item searches may improve the user experience.

Algorithms for extracting image features based on deep convolutional neural network (CNN) [7, 8] and approximate nearest neighbor (ANN) search [2, 4] can be used to realize a simple visual search system. However, even if these simple systems can retrieve visually similar images, their results could be nonoptimal. C2C e-commerce site search algorithms tend to extract items listed by professional sellers even if more relevant items are listed by nonprofessional sellers because the query images are often more visually similar to images taken by professionals than those provided by nonprofessionals, especially in apparel categories. Specifically, fitted apparel images (Figure 1b) are likely to be retrieved in response to a fitted apparel query image (Figure 1a). In this paper, we call apparel "fitted" if it is pictured being worn by a model and "flat" if it is instead laid flat on a surface. Professional and nonprofessional sellers tend to upload fitted and flat apparel images, respectively. Searches that return many items listed by professional sellers can cause problems for C2C e-commerce sites, for example, by hurting buyer experience and discouraging nonprofessional sellers from listing items [1].



(a) Query Image (b) Baseline (c) Proposed

Figure 1: Query image and its visual search results among 100 million items.

To manage these issues so as to retrieve more flat apparel items, we developed an image representation technique that closes the visual gap between fitted apparel query images and flat apparel images in a database. The technique consists of extracting features using a lightweight deep CNN and transforming query features; it enables the retrieval of flat apparel images (Figure 1c) from a fitted apparel query image (Figure 1a). Moreover, the feature transformation step can be applied to any query vector because it causes no significant side effects to flat apparel and nonapparel query vectors. Thus, additional information of whether the query image contains fitted apparel is not required before feature transformation. Our experiments demonstrate that more flat apparel images are correctly discovered through query feature transformation for a fitted apparel query image without serious negative impacts on flat apparel and nonapparel query images.

In: J. Degenhardt, S. Kallumadi, U. Porwal, A. Trotman (eds.):
Proceedings of the SIGIR 2019 eCom workshop, July 2019, Paris, France, published at http://ceur-ws.org

2 RELATED WORK

Some e-commerce sites, such as Alibaba and eBay, have introduced visual search systems that enable users to search for products using images [9, 11]. These systems are basically composed of deep CNN-based feature extraction and a nearest neighbor search. A method of discovering more items relevant to a query image involves the training of a deep CNN model with a triplet loss; however, building and updating a dataset for training such models is infeasible for a massive and volatile inventory marketplace. Although implicit feedback, such as page views and click logs, allows for model training with a triplet loss even in such a marketplace [11], implicit feedback is available only after launching a visual search system into production. This paper proposes an image representation method with query feature transformation; this method closes the gap between a fitted apparel query vector and flat apparel database vectors without time-consuming human relevance assessments.

3 VISUAL SEARCH ARCHITECTURE

The proposed visual search architecture simply consists of image feature extraction, query feature transformation, and a nearest neighbor vector search. For C2C e-commerce sites specifically, this feature transformation closes the distance between a fitted apparel query vector and flat apparel database vectors. An approximate nearest neighbor (ANN) algorithm accomplishes the nearest neighbor search in a large database within a practical runtime.

3.1 Image Representation

3.1.1 Feature Extraction Model. For feature extraction, we adopted MobileNetV2 [6], which is a state-of-the-art lightweight CNN model. Sending query images at a large scale to an e-commerce visual search system from user devices can cause network traffic problems. One solution to this issue is edge computing, through which image features are extracted on an edge device or a smart device. Such a lightweight extraction model works efficiently in an edge device and consumes only several megabytes of memory space.

We prepared a dataset consisting of images and their metadata collected from an online C2C marketplace with over one billion listings. The dataset has 9 million images belonging to 14,000 classes, which are combinations of item brands, textures, and categories—for example, Nike striped men's golf polo. Images from nonapparel categories, such as laptops, bikes, and toys, are included in the dataset.

One of the model's hyper parameters is a width multiplier [6]; for a given layer and width multiplier α, the number of output channels N becomes αN. The model was trained on the dataset with a width multiplier of 1.4. The output of the global average pooling layer was used as an image feature vector that has $1,792$ $(1,280 \times 1.4)$ dimensions. Then, the feature vectors of the query and database images were extracted using the same feature extractor.

3.1.2 Query Feature Transformation. Only the query feature vectors were calibrated using a feature transformation vector, which expresses a human feature vector intuitively, to close the gap between fitted apparel query feature vector and flat apparel database feature vectors. The feature transformation vector was trained through Algorithm 1 with 80,040 images (54,145 of fitted apparel and 25,895 of flat apparel) belonging to 15 apparel categories, such as tops, jackets, pants, and hats. In the training step, a gap vector, which represents the difference between fitted and flat apparel feature vectors, was calculated for each category and the feature transformation vector was computed by averaging the gap vectors. For a query, the transformation simply subtracts the feature transformation vector from a query image feature vector (Algorithm 2); its computation time is negligibly small.

Algorithm 1 Generate Feature Transformation Vector

Input:
$\boldsymbol{h}_{\{1,\ldots,C\},\{1,\ldots,N_c\}}$, # Feature vectors of fitted apparel images
$\boldsymbol{l}_{\{1,\ldots,C\},\{1,\ldots,M_c\}}$ # Feature vectors of flat apparel images
$\boldsymbol{h}_{c,n}$ and $\boldsymbol{l}_{c,m}$ represent the n-th and m-th feature vectors of category c, respectively.
Output:
Feature transformation vector $\hat{\boldsymbol{z}}$

for $c \leftarrow 1$ to C **do**
  $\bar{\boldsymbol{h}}_c \leftarrow \text{Median}(\boldsymbol{h}_{c,1}, \ldots, \boldsymbol{h}_{c,N_c})$ # Fitted apparel vector (Median vector of the fitted apparel vectors)
  $\bar{\boldsymbol{l}}_c \leftarrow \text{Median}(\boldsymbol{l}_{c,1}, \ldots, \boldsymbol{l}_{c,M_c})$ # Flat apparel vector (Median vector of the flat apparel vectors)
  $\boldsymbol{t}_c \leftarrow \bar{\boldsymbol{h}}_c - \bar{\boldsymbol{l}}_c$ # Subtracting the flat apparel vector from the fitted apparel vector.
  $\boldsymbol{z}_c \leftarrow \text{Maximum}(\boldsymbol{t}_c, \boldsymbol{0})$ # Replacing negative elements with 0
  $\hat{\boldsymbol{z}}_c \leftarrow \frac{\boldsymbol{z}_c}{\|\boldsymbol{z}_c\|_2}$ # L2 Normalization. $\hat{\boldsymbol{z}}_c$ is a gap vector of category c
$\boldsymbol{z} \leftarrow \text{Average}(\hat{\boldsymbol{z}}_1, \ldots, \hat{\boldsymbol{z}}_C)$ # Averaging the gap vectors
$\hat{\boldsymbol{z}} \leftarrow \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_2}$ # L2 Normalization
return $\hat{\boldsymbol{z}}$ # Feature transformation vector

Algorithm 2 Feature Transformation

Input:
$\boldsymbol{q}$, # Feature vector of the query image
$\hat{\boldsymbol{z}}$ # Feature transformation vector
Output:
Transformed query vector $\hat{\boldsymbol{p}}$

$\hat{\boldsymbol{q}} \leftarrow \frac{\boldsymbol{q}}{\|\boldsymbol{q}\|_2}$ # L2 Normalization
$\boldsymbol{t} \leftarrow \hat{\boldsymbol{q}} - \hat{\boldsymbol{z}}$ # Subtracting feature transformation vector from query vector
$\boldsymbol{p} \leftarrow \text{Maximum}(\boldsymbol{t}, \boldsymbol{0})$ # Replacing negative elements with 0
$\hat{\boldsymbol{p}} \leftarrow \frac{\boldsymbol{p}}{\|\boldsymbol{p}\|_2}$ # L2 Normalization
return $\hat{\boldsymbol{p}}$ # Transformed query vector

The feature vector extracted from MobileNetV2 initially lacks negative value elements owing to the use of the rectified linear unit (ReLU) activation function [5]. The negative value elements in the feature vector space can be treated as unnecessary, that is, elements are replaced with zero in Algorithms 1 and 2, a step that is key to preventing side effects in query feature transformation. Even if the feature transformation designed for a fitted apparel query vector is applied to a flat apparel or nonapparel query vector, the essential feature is still preserved by removing negative value elements.

3.2 Nearest Neighbor Search

In large-scale e-commerce, ANN searches outperform brute force in finding the nearest neighbors of a transformed query vector from the database vectors. ANN algorithms, such as IVFADC [2] and

Table 1: Visual Search Results for Apparel Categories (*mAP*@100)

	Flat Apparel		Fitted Apparel		Fitted Apparel (Cropped)	
	Baseline	Proposed	Baseline	Proposed	Baseline	Proposed
T-Shirts	0.844	**0.895**	0.004	**0.376**	0.042	**0.542**
Sweaters	0.926	**0.967**	0.002	**0.456**	0.053	**0.670**
Hoodies	0.942	**0.977**	0.053	**0.691**	0.211	**0.756**
Denim Jackets	0.982	**0.993**	0.004	**0.778**	0.041	**0.850**
Down Jackets	0.972	**0.995**	0.115	**0.815**	0.313	**0.866**
Jeans	**0.878**	0.822	0.001	**0.381**	0.095	**0.737**
Casual Pants	0.889	**0.933**	0.002	**0.475**	0.139	**0.690**
Knee-Lengh Skirts	0.718	**0.732**	0.000	**0.081**	0.090	**0.257**
Long Skirts	0.567	**0.614**	0.004	**0.180**	0.041	**0.244**
Dresses	0.847	**0.922**	0.001	**0.226**	0.018	**0.254**

Rii [4], allow us to retrieve the nearest neighbors in a practical runtime. In our experiments, we used IVFADC to retrieve visually similar images from among 100 million images.

4 EXPERIMENTS

We conducted experiments to evaluate the proposed method. For these experiments, we collected 20,000 images from a C2C marketplace: half of these images were those of flat apparel and the remaining were fitted apparel images. The flat apparel images belong to ten categories, shown in the first column of Table 1. Fitted apparel images not belonging to the ten categories, such as jerseys and polo shirts, were also included. From the 20,000 images, 2,000 were used as query images, from among which 100 images were randomly selected from the 10 categories for each flat and fitted apparel class. The remaining 18,000 images were treated as database images. For fitted apparel queries, cropped images of the query objects were prepared manually from the original images to reduce the influence of the background.

The mean average precision at 100 (*mAP*@100), defined as follows, was used as an evaluation measure for each category.

$$mAP@K = \frac{\sum_{q=1}^{N} AP@K(q)}{N},$$

where

$$AP@K = \frac{\sum_{k=1}^{K} (P@k \cdot I(k))}{\sum_{k=1}^{K} I(k)}, \quad P@k = \frac{\sum_{n=1}^{k} I(n)}{k},$$

$$I(i) = \begin{cases} 1 & i\text{-th item is flat apparel in the same category as the query} \\ 0 & \text{otherwise} \end{cases},$$

N is the number of query images, and $AP@K$ and $P@k$ indicate the average precision at K and precision at k for each query, respectively. A retrieved item is recognized as correctly selected only when it is an image of flat apparel in the same category as the query.

For baseline image representation, a vector from the global average pooling layer of MobineNetV2, described in Section 3.1, was used for query and database images. Our proposed method also uses the same feature extractor and transforms query vectors. Because the number of database images used in this experiment was relatively small, the nearest vectors were greedily retrieved using cosine similarity. Table 1 compares the *mAP*@100 of the baseline and proposed image representations for flat and fitted apparel query images. The results demonstrate a significant improvement for the fitted apparel queries in every category. Although query feature transformation was designed to close the gap between fitted and flat apparel vectors, it also positively influenced flat apparel queries. These results imply that our proposed method enables more essential features to be extracted from query images.

We also collected 100 million images belonging to over 1000 item categories, including nonapparel images. For such large-scale data, ANN algorithms allow us to retrieve the nearest neighbors within a practical runtime. Figure 2 presents the visual search results with IVFADC [2] (code length per vector: 64 bytes, number of cells: 8,192, number of cells visited for each query: 64), from the 100 million images for fitted apparel and nonapparel queries. To demonstrate the versatility of the proposed method, the fitted apparel queries also contain images from a different dataset, ATR [3], which was originally used for a human parsing task. For fitted apparel queries, our proposed method retrieved a greater number of visually similar flat apparel items (1st–8th rows in Figure 2). In addition, no serious negative impact was observed for nonapparel queries: visually similar items to the query images were successfully extracted (9th–13th rows in Figure 2). The runtimes of the image feature extraction method and the nearest-100 vector search were approximately 40 and 70 ms, respectively, using an 8-core 2.3 GHz CPU. By simultaneously processing multiple query images and/or using GPUs, the runtime per query could be made faster.

5 CONCLUSION

This paper proposed an image representation technique for visual search at C2C e-commerce sites. The proposed method, comprising a deep CNN-based feature extraction and query feature transformation, significantly improves conventional visual search methods for comparing images of fitted and flat apparel. Additionally, the proposed method did not negatively impact either flat apparel or nonapparel queries in a serious manner. The performance and total runtimes of our visual search system were practical in the experiments described, indicating that it can be successfully deployed to a major online C2C marketplace. After the system is widely used in production, further improvement is expected using real query images and implicit feedback.

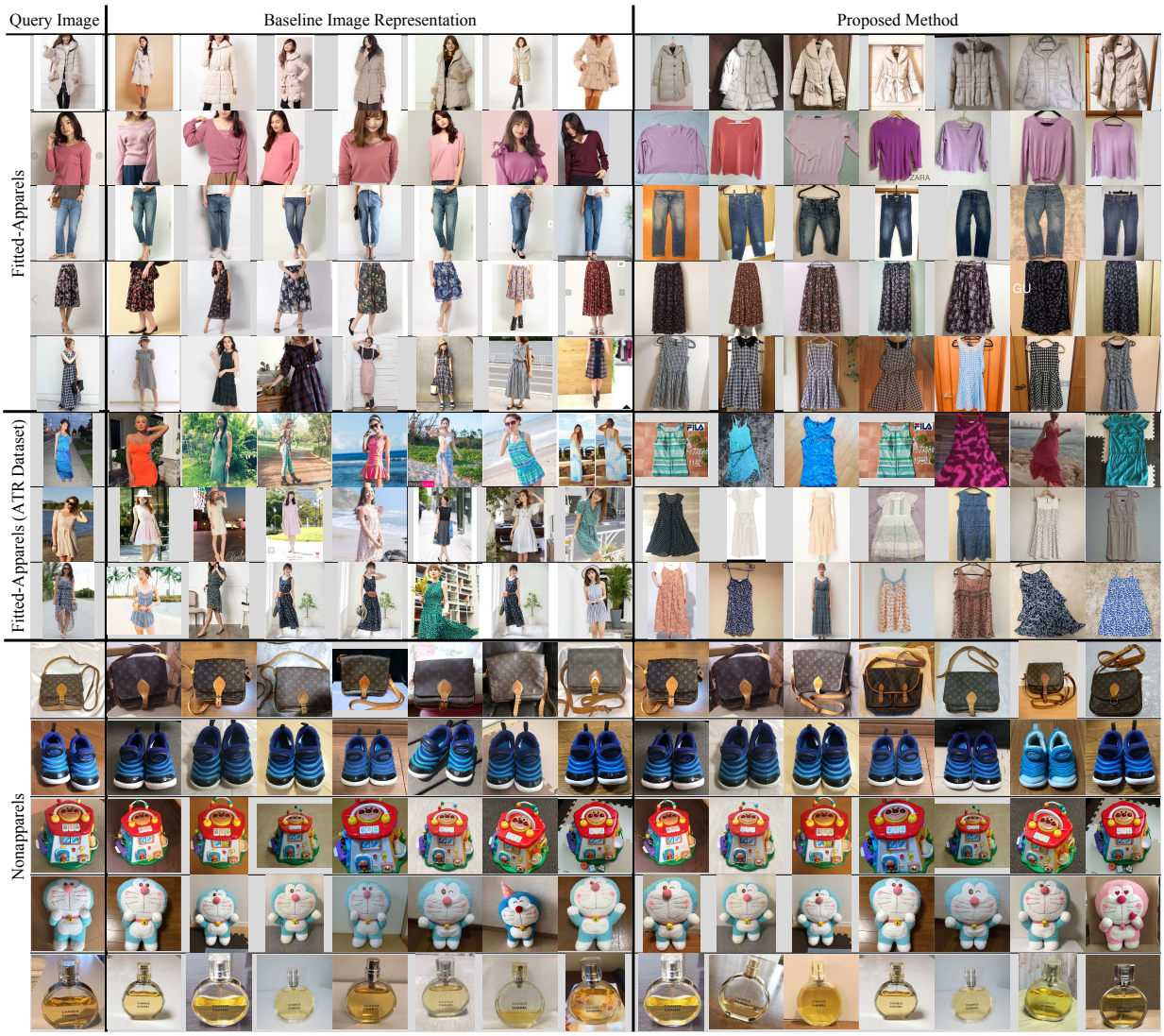


Figure 2: Visual search results from 100 million images. The first column shows query images and the next seven columns show the results with the baseline image representation. The remaining columns show the results obtained using query feature transformation. Our method successfully retrieved more flat apparel images corresponding to the fitted apparel queries without negatively impacting nonapparel queries.

REFERENCES

[1] A. Hagiu and R. Simon. 2016. Network Effects Aren't Enough. *Harvard Business Review* 94, 4 (April 2016), 65–71.

[2] H. Jegou, M. Douze, and C. Schmid. 2011. Product Quantization for Nearest Neighbor Search. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 1 (Jan. 2011), 117–128.

[3] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan. 2015. Deep Human Parsing with Active Template Regression. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 12 (Dec. 2015), 2402–2414.

[4] Y. Matsui, R. Hinami, and S. Satoh. 2018. Reconfigurable Inverted Index. In *Proceedings of the 26th ACM International Conference on Multimedia (MM '18)*. 1715–1723.

[5] V. Nair and G. E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML'10)*. 807–814.

[6] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR '18)*.

[7] X. Song, S. Jiang, and L. Herranz. 2017. Multi-Scale Multi-Feature Context Modeling for Scene Recognition in the Semantic Manifold. *Trans. Img. Proc.* 26, 6 (June 2017), 2721–2735.

[8] A. Babenko Yandex and V. Lempitsky. 2015. Aggregating Local Deep Features for Image Retrieval. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV '15)*. 1269–1277.

[9] F. Yang, A. Kale, Y. Bubnov, L. Stein, Q. Wang, H. Kiapour, and R. Piramuthu. 2017. Visual Search at eBay. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. 2101–2110.

[10] A. Zhai, D. Kislyuk, Y. Jing, M. Feng, E. Tzeng, J. Donahue, Y. L. Du, and T. Darrell. 2017. Visual Discovery at Pinterest. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion)*. 515–524.

[11] Y. Zhang, P. Pan, Y. Zheng, K. Zhao, Y. Zhang, X. Ren, and R. Jin. 2018. Visual Search at Alibaba. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '18)*. 993–1001.