Ten-million-order Human Database for World-wide Fashion Culture Analysis

Hirokatsu Kataoka, Yutaka Satoh
National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki, Japan
{hirokatsu.kataoka, yu.satou}@aist.go.jp

Kaori Abe, Munetaka Minoguchi, Akio Nakamura
Tokyo Denki University (TDU)
Senju Asahi-cho, Adachi-ku, Tokyo
minoguchi.m@is.fr.dendai.ac.jp

Abstract

The paper proposes a huge geo-tagged image database, referred to as the Fashion Culture DataBase (FCDB), which is considered to be a semi-automatic image collection in cloud-based services such as social network services (SNS). In the stage of database construction, we introduce the large-scale data collection, refinement, and representation for SNS-based analysis. The proposed FCDB consists of 25,707,690 images for use in (i) inter-city similarity analysis and (ii) fashion trends visualization. We also present a simple but effective representation for the spatial and temporal analysis. Finally, we visualize an inter-city graph and the yearly trend change with the refined FCDB.

1. Introduction

The usage of Social Network Services (SNS) has been increasing in the last decade [8]. As of 2017, a large amount of images are daily posted to photo sharing websites such as Facebook, Instagram and Flickr. The huge amount of images are taken by users positioned by various places. Therefore, it can be said that a user's preference and tendency are reflected in the images. We believe that the kind of sharing images in the world must be recognized and analyzed by the latest vision-based techniques.

Conventionally, regional analysis have been performed using geo-tagged images on cloud-based service [2, 3, 6, 7, 10, 17]. Zhou *et al.* attempted to implement scene-based urban recognition focusing on the cityscape of 21 cities representing cultures using tagged images from Google Street View, Flickr, and Panoramio (this is the city perception) [17]. Simo-Serra *et al.* proposed a Fashion 144k dataset that includes tagged fashion snapshots and user ratings as its coordinates.

Figure 1. FCDB (top) FCDBv2 (bottom), containing a large number fashion images. FCDB(v1) contains roughly collected images but FCDBv2 have eliminated noisy images by a simple modification. Although it is desirable that a whole body image is clipped, FCDB picks up the image of noise (e.g., background, human-like pillar and sculpture) and upper body (only). Our FCDBv2 succeeded to collect ideal images most of the cases. We randomly sampled all images which are not curated in the figure.

Fashion144k is possible to analyze the correlation between Fashionability of country and GDP [13]. Moreover, StyleNet [14] which learned in the framework of weak supervision in the Fashion144k dataset can extract fashion

Table 1. Fashion-oriented databases. The listed databases are limited to closely related or representative one. Our FCDBv2 is the most biggest database in terms of #images.

Database	Task	# images	# category	Geo-tag	Person-tag	time-stamps
HipsterWars [9]	Style classification	1,893	5			
Fashionista [16]	Parsing, pose estimation	158,235	53		✓	
Fashion144k [13]	Style classification	144,169	N/A	✓		
FashionStyle14 [15]	Style classification	13,126	14			
DeepFashion [11]	Attribute estimation	800,000	1,050		✓	
FCDBv2 (ours)	Fashion trends analysis	**25,707,690**	16	✓	✓	✓

style from image which has noisy background. Therefore, we believe that a very large-scale data collection by cloud-based service allows us to analyze a world-wide fashion trends by combining a sophisticated fashion-oriented descriptor.

In the present paper, we propose a Fashion Culture DataBase (FCDB), which contains 25M images on Flickr, a cloud-based service. The FCDB is considered to be a semi-automatic image collection and refinement tool with a simple detector and classifier. The database enables analysis of spatial and temporal fashion attributes. In the spatial analysis, we attempt to visualize inter-city similarity of 16 cities based on a previous study [17]. In the temporal analysis, we analyze temporal subtraction between two consecutive times (e.g., years y & $y+1$) in order to delineate an increased / decreased fashion trends at the moment.

The following are the contributions of the present study.

(i) We propose a huge-scale FCDB that contributes to spatio-temporal and wide-range clothing image analysis on cloud service. The FCDB contains noisy images (see Figure 1) depending on failure detection. We simultaneously consider a data refinement for the fashion-oriented database and present a simple trick that is sufficient for refining the FCDB. Hereafter, the refined FCDB is referred to as *FCDBv2*.

(ii) Using FCDBv2, we analyze and visualize an inter-city similarity graph and yearly fashion trends. We also evaluate the difference between roughly collected (FCDB) and refined (FCDBv2) databases. For example, we show the effectiveness of noise canceling in Figure 1.

2. Fashion Culture DataBase (FCDB)

2.1. Overview

The FCDB was collected from the Yahoo! Creative Commons 100 M Database (YFCC100M) [1], which contains 100 million Flickr images. We focus on 21 global cities based on city perception [17]. However, we exclude the cosmopolitan cities if the number of collected images is less than 100K. Consequently, 16 of the 21 cosmopolitan cities remain. In order to perform temporal fashion analysis as fashion trend, we insert a time-stamp for each image from 2000 to 2015 into the database. We also crop to person-centered patches with the Faster R-CNN [12] (see Figure 2). We generated FCDBv2, which is a large fashion-oriented database that includes 25,707,690 clothing images by refining the FCDB. To the best of our knowledge, this is the largest existing fashion-oriented database (see Table 1) by considering image noise in a database.

2.2. Detailed collection and refinement

As shown in Figure 2, the FCDBv2 contains (i) captured images from the YFCC100M dataset, (ii) person-cropped images (we treat clothing images), and (iii) geo-location and time-stamp information. We set a 100-km radius around a city in order to collect images based on geo-location.

The 16 cities on the FCDB are {*London, New York, Boston, Paris, Toronto, Barcelona, Tokyo, San Francisco, Hong Kong, Zurich, Seoul, Beijing, Bangkok, Singapore, Kuala Lumpur, and New Delhi*}. We considered that the areas do not overlap each other. To create the images in (ii), we apply the Pascal VOC [5] pre-trained VGG16 model for the Faster R-CNN [12]. We experimentally set the thresholding value as 0.8 and use only the person-label. A set of data consisting of a geo-tag and a time-stamp in (iii) is replicated from the YFCC100M dataset. In the first execution, we have collected 76,532,519 images with semi-automatic image collection and annotation. However, the fashion-oriented database contains noisy images, such as partially cropped persons (e.g. face only) and backgrounds (e.g. tree, traffic sign): therefore, we describe how to refine the FCDB to obtain FCDBv2 in the next subsection.

We consider how to exclude noisy images that are included in the roughly collected FCDB. The refinement strategy is to scan all images with a simple classifier by combined StyleNet [14] and an kernel SVM. In order to collect a sophisticated fashion-oriented database, we treat the FCDB refinement as binary classification between "street-fashion-snap-like whole body" and "other cropped images such as partial body or backgrounds without a person". We trained and refined the database with 2,886 carefully annotated objective images and a large number of randomly selected negative images. The effectiveness of data refine-

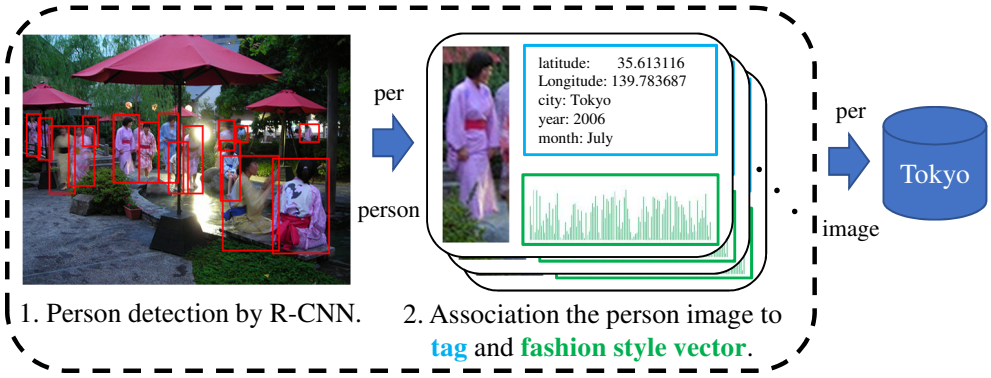


Figure 2. Construction of the FCDB: We operate person cropping by (Faster) R-CNN [12], adding geo-location (latitude/longitude/city), and inserting time-stamp. We associated a person photograph with tag information such as geo-location and time-stamp and fashion style vector with StyleNet [14]

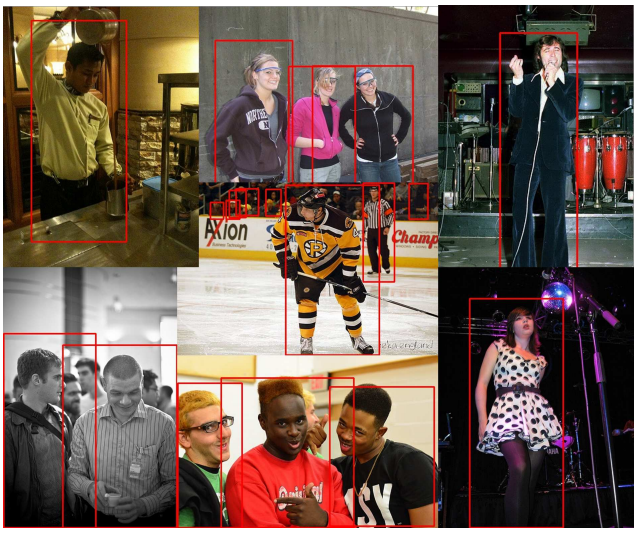


Figure 3. Automatically annotated bounding boxes with Faster R-CNN and binary SVM classifier. We firstly picked the person's bounding boxes with Faster R-CNN. To eliminate wrong boxes, we refine with binary classifier whether person or not. The binary classifier is based on SVM with fashion snaps.

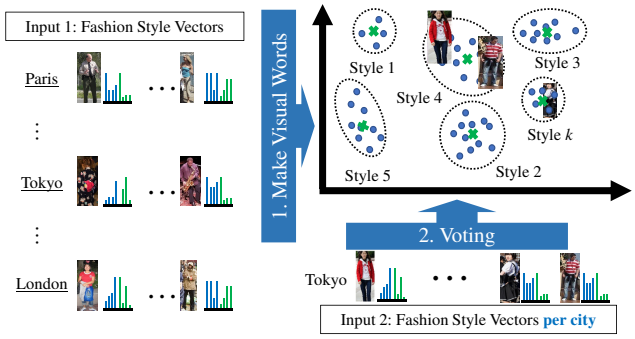


Figure 4. Fashion Style Distribution (FSD) for per-city feature representation. FSD is made by StyleNet and bag-of-words (BoW) at each city.

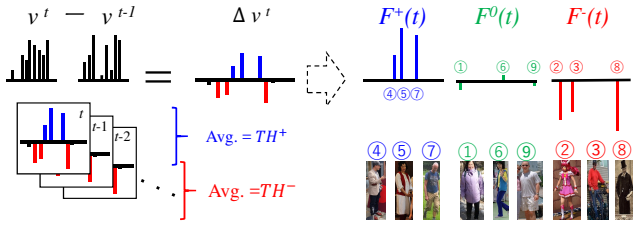


Figure 5. Fashion Trend Descriptor (FTD) for temporal trend analyzer. Two consecutive FSD are subtracted to disclose a fashion trend.

ment is shown in the experimental section.

Fashion Trend Descriptor (FTD). In order to clarify the fashion trend, which explains the temporal change, we propose the Fashion Trend Descriptor (FTD) (see Figure 5). The FTD, F^t, which is described as a labeled histogram on time t, the degree of which is Δv_i^t for fashion style i, is calculated by subtracting FSDs of pseudo fashion style i between consecutive times $t-1$ and t. Classes are then distinguished and inserted into three categories: increased F^+, unchanged F^0, and decreased F^- fashion styles. As a result, which fashion styles appeared/disappeared or how much they translated by analyzing FTD.

3. Analyzing methods

Fashion Style Distribution (FSD). We consider that fashion trends appear as tendencies of the distribution of the appearance frequency of fashion styles. Thus, we propose a Fashion Style Distribution (FSD), which is represented by StyleNet [14] bag-of-words (BoW) [4]. The framework is shown in Figure 4. We perform k-means clustering to define fashion style as clusters automatically calculated from 100,000 randomly selected clothing images. We treat the cluster as *pseudo* fashion style. Although the pseudo fash-

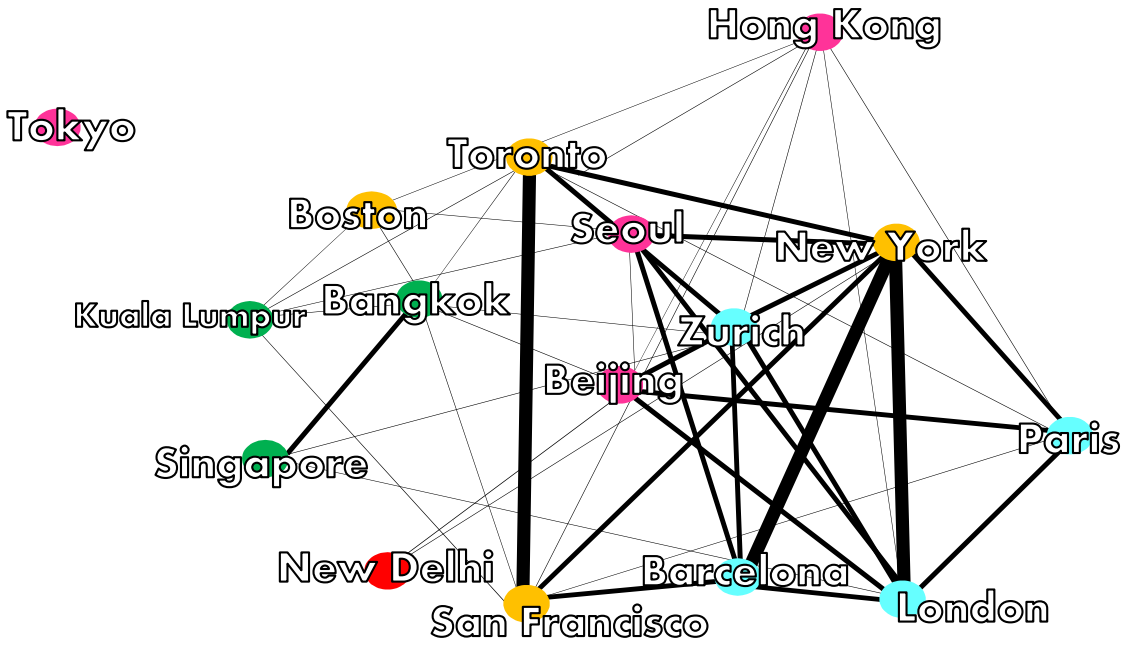
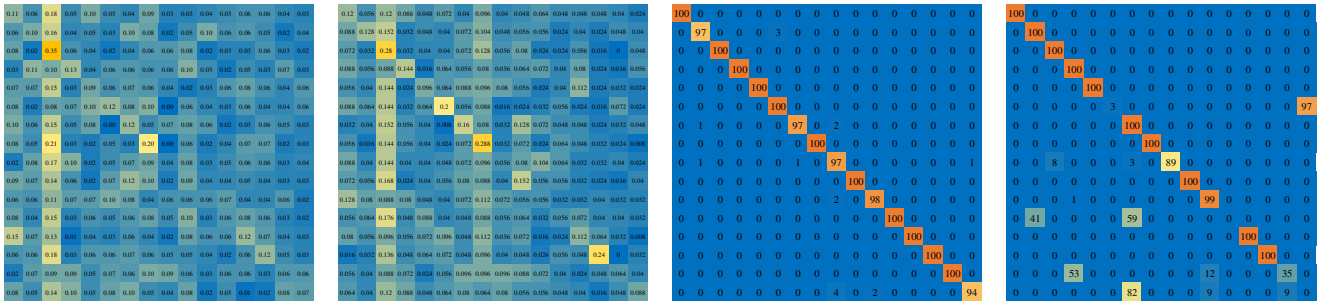


Figure 6. Fashion-based city similarity graph



(a) *accuracy* = 11.5 % on FCDB (b) *accuracy* = 14.3 % on FCDBv2 (c) *accuracy* = 98.9 % on FCDB (d) *accuracy* = 83.8 % on FCDBv2

Figure 7. (a), (b) : Results of fashion-based city perception. (c), (d) : Results of FSD-based city perception: The order of the labels is as follows. {Bangkok, Barcelona, Beijing, Boston, Hong Kong, Kuala Lumpur, London, New Delhi, New York, Paris, San Francisco, Seoul, Singapore, Tokyo, Toronto, Zurich}

ion style is not a real fashion style, we can confirm the improvement from FCDB to FCDBv2 with FSD (see Figure 6 in city perception with FCDBv2, Figure 7 in FSD-based city perception and Figure 8 in fashion trend analysis). What is better, the FSD is created without any supervision. Here, we experimentally set the cluster size k to be 1,000[1]. And these 1,000 visual words express *pseudo* fashion styles. By calculating the FSD in 16 cities, it is possible to histogramize the appearance distribution of fashion style at each city. By using the FSD, we directly calculate inter-city similarity graph (see Figure 6).

[1]Exhaustive tuning is better way, however, our database contains over 25M images to investigate various parameters. We experimentally tried some parameters to decide the configuration.

4. Evaluation

We verify the geo-spatial analysis based on a previous study [17], but the temporal analysis is original setup.

We discuss the effectiveness of data refinement after the experiments in this section.

4.1. Geo-spatial fashion analysis

16 city perception. Figure 7 (a) and (b) describe per-image-level city perception with StyleNet + SVM following to Zhou *et al.* [17]. We show that our FCDBv2 is +2.8 overall accuracy (14.3 vs. 11.5) better than roughly collected FCDB. This is pure improvement with dataset refinement.

Figure 7 (d) shows the confusion matrix of the city o with FSD, a collective image analysis. A vector FSD was made with 10,000 randomly selected images, and we pre-

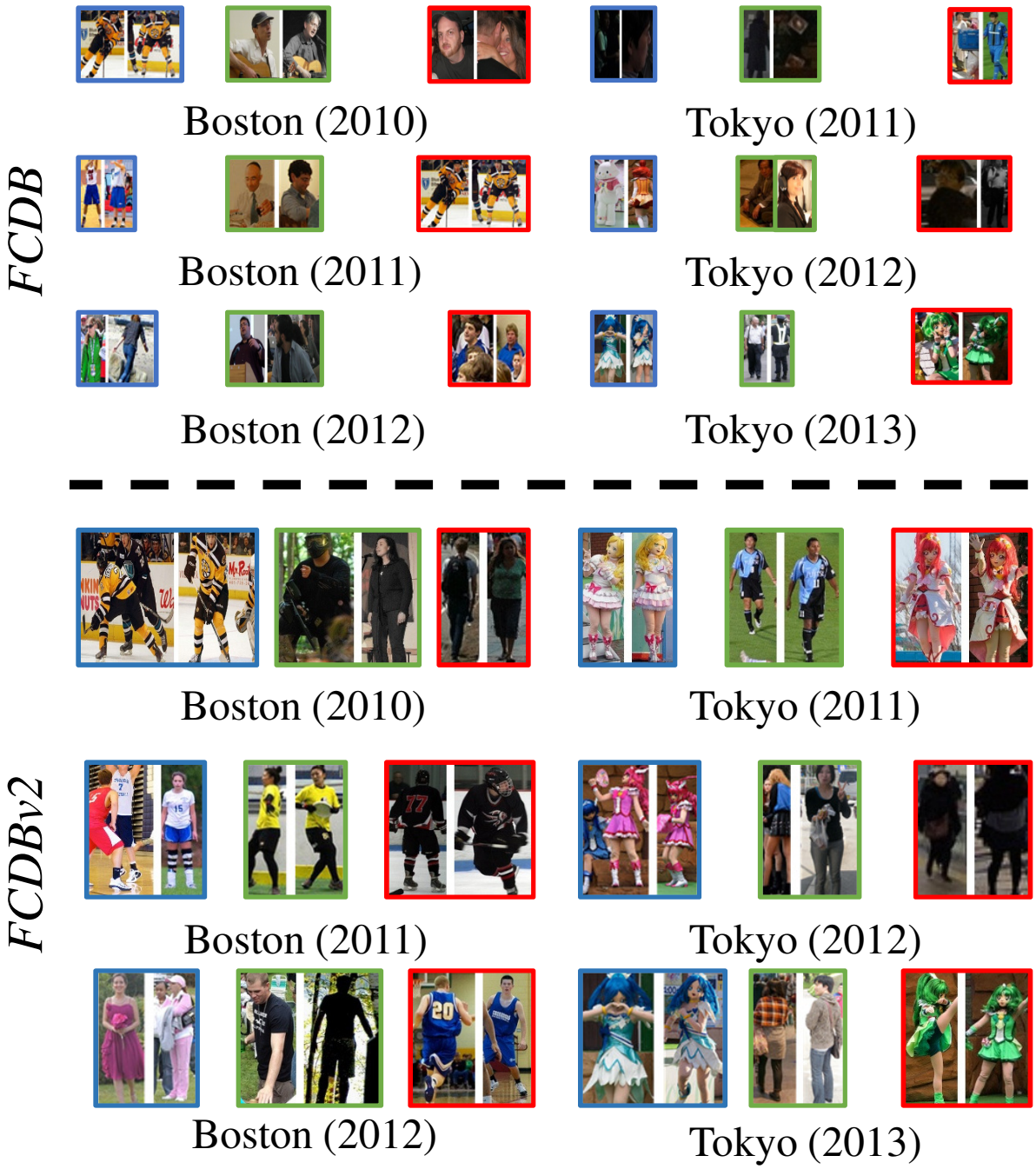


Figure 8. Visualization of fashion trends (top: FCDB, bottom: FCDBv2): The figure shows randomly selected cutting-edge of each styles per city and year (increased: blue F^{+}, zero-mean: green F^{0}, decreased: red F^{-}).

pared {500, 100} FSD at each city for {training, testing}. The overall performance rate of city perception with the FSD is 83.8%. The result suggests that the FSD emphasizes the city characteristics by comparing with the a single StyleNet vector. We also confirmed that fashion tends to be a more discriminative feature (e.g., New York is more similar to other Western countries than to Asian countries). The results indicate that FSD is more discriminative than the scene-based feature in [17].

16 city perception. We show the inter-city similarity graph among 16 cities with FSD. Figure 6 illustrates the fashion-based city similarity graph. According to Figure 6 and 7 (d), the cities which have perfect (100%) perception (e.g., Paris, Tokyo, Boston, Hong Kong, and New Delhi)

tend to have separable characteristics, such as cultural background and climate. On one hand, European and North American countries (e.g. London, Zurich, San Francisco, and Toronto) are not perfect perception. Therefore, the similarity of fashion trends comes from similar characteristics, such as language, climate, and geographical location.

4.2. Temporal fashion analysis

Here, the Figure 8 shows a fashion trend analysis for consecutive years. We assign the FTD to explicitly depict fashion trends as a result of temporal subtraction. The FTD is executed performed with images over 16 years (from 2000 to 2015) with FCDB and FCDBv2. In particular, the parameters F^+ and F^- visualize the changes in fashionability according to differences in the FSD for each consecutive year. Figure 8 illustrates the changing fashion trends for the randomly selected cities and years[2].

Figure 8 shows samples of the F^+ (blue), F^0 (green), and F^- (red) which are salient in an arbitrary year / city. Six configurations of city (year) (e.g., Boston (2010)) are listed by comparing between FCDB and FCDBv2. FCDBv2 captures the temporal fashionability change purely than FCDB by refinement. In Boston (2010), a sports uniform appears as a trend that frequently occurred. The appearance of a sports uniform suggested that people in Boston are interested in sports. (The image for Boston (2010) appears to be an ice hockey uniform.) According to our survey, the local Boston ice hockey team acquired prominent titles in 2010, after not receiving a title for a long period of time. However, the sports uniform disappeared as a trend in Boston (2011). In Tokyo (2011–2013), costumes of animated characters continued to trend, which is a frequent occurrence each year. Subcultures such as Japanese anime and manga are popular and thus tended to appear in images. The FCDBv2 in Tokyo contains a large number of costume images taken by users. We believe that the reason why Tokyo is distinguished from other countries shown in Figure 6.

Given these visualizations, it was suggested that, by combining FCDBv2 and FTD, it is possible to evaluate spatio-temporal properties of fashion trends appearing in cloud-based service.

4.3. Discussion

We discuss primarily the effectiveness of data refinement in terms of the FCDB and FCDBv2.

Per-image city perception (Figure 7 (a) and (b)). We verified the per-image city perception with StyleNet and SVM. We randomly selected {500, 125} images at each city for {training, testing}. The refinement made a +2.9% improvement in overall accuracy; therefore the uniqueness among cities was improved.

FSD-based city perception (Figure 7 (c) and (d)). In the FSD-based city perception, the refinement database made a 15.1% lower accuracy. The FSD with roughly collected FCDB is unexpectedly higher accuracy than FCDBv2 in Figure 7 (c). The roughly collected FCDB has noisy data as shown in Figure 1. A biased data tends to contain many additional images like landscapes to unexpectedly improve the accuracy. The 76M images on FCDB seems to have a lot of noise but our FCDBv2 decrease the noise with a simple classification.

Inter-city similarity graph (Figure 6) and trend analysis (Figure 8). The FCDBv2 allows successful visualization of inter-city similarity and trend analysis. In particular, in the trend analysis, the effectiveness of refinement is remarkably. The FCDB includes non-fashion snapshots, such as the face-only and upper-body images on the left-hand side of Figure 8. Unlike the FCDB, the FCDBv2 uses only whole-body images from among randomly selected samples for the trend analysis. Since clothes are closely depending on a daily living, fashion trends are partially formed by the history and climate of the area (see Figure 8). In our geospatial and temporal analysis on FCDBv2, we could discuss regionality and temporal change of fashion trends on cloud-based service where variety of subjects in photo can be seen. In our insight, it is suggested that FCDBv2 can reproduce real-world fashion trends.

5. Conclusion

We proposed a huge fashion-oriented database, named the Fashion Culture DataBase (FCDB) which is capable to analyze fashion trends in SNS. FCDB has 25,707,690 geotagged and time stamped images in 16 cosmopolitan cities. In the analysis, to visualize the cutting-edge fashion trends, we also proposed a fashion trends descriptor (FTD) that is composed by a fashion style descriptor, a codeword vector, and temporal subtraction. In addition, we refined the FCDB for world-range fashion trends analysis. The combination of our FCDB and FTD significantly visualizes world-wide fashion trends in time series. In the future, we plan to collect fashion snaps over a longer timeframe (e.g., 30 years) in order to observe the periodic changes of fashion cultures in different countries.

References

[1] T. Bart, D. A. Shamma, F. Gerald, E. Benjamin, N. Karl, P. Douglas, B. Damian, and L. Li-Jia. YFCC100M: The new data in multimedia research. *Commun. ACM*, 59:64–73, jan 2016.

[2] D. Carl, S. Saurabh, G. Abhinav, S. Josef, and A. A. Efros. What makes paris look like paris? *ACM Trans. Graph.*, 31:101:1–101:9, jul 2012.

[2] We took a couple of representative images that are the nearest images to a randomly selected centroid. The nearest images represent a strong fashionability at the codeword.

[3] D. J. Crandall, L. Backstrom, D. Huttenlocher, and K. Jon. Mapping the world's photos. In *Proceedings of the 18th International Conference on World Wide Web*, pages 761–770, New York, NY, USA, 2009.

[4] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[5] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, jan 2015.

[6] J. Hays and A. A. Efros. IM2GPS: estimating geographic information from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[7] J. Heinly, J. L. Schonberger, E. Dunn, and J. Frahm. Reconstructing the world* in six days *(as captured by the yahoo 100 million image dataset). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[8] S. Inc. statista: Number of monthly active facebook users worldwide as of 3rd quarter 2017 (in millions), 2017. `https://goo.gl/pFyiUh`.

[9] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In *European Conference on Computer Vision (ECCV)*, 2014.

[10] L. Liu, B. Zhou, J. Zhao, and B. D. Ryan. C-image: city cognitive mapping through geo-tagged photos. *Geo Jounal*, 81:817–861, dec 2016.

[11] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2016.

[12] R. Shaoqinga, H. Kaiming, G. Ross, and S. Jian. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc., New York, NY, USA, 2015.

[13] E. Simo-Serra, S. Fidler, F. M. Noguer, and R. Urtasun. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[14] E. Simo-Serra and H. Ishikawa. Fashion style in 128 floats:joint ranking and classification using weak data for feature extraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] M. Takagi, E. Simo-Serra, S. Iizuka, and H. Ishikawa. What makes a style: Experimental analysis of fashion prediction. In *Proceedings of the International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[16] K. Yamaguchi, M. H. Kiapour, and T.L.Berg. Parsing clothing in fashion photographs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[17] B. Zhou, L. Liu, A. Oliva, and A. Torralba. Recognizing city identity via attribute analysis of geo-tagged images. *ECCV: European Conference on Computer Vision*, pages 519–534, 2014.

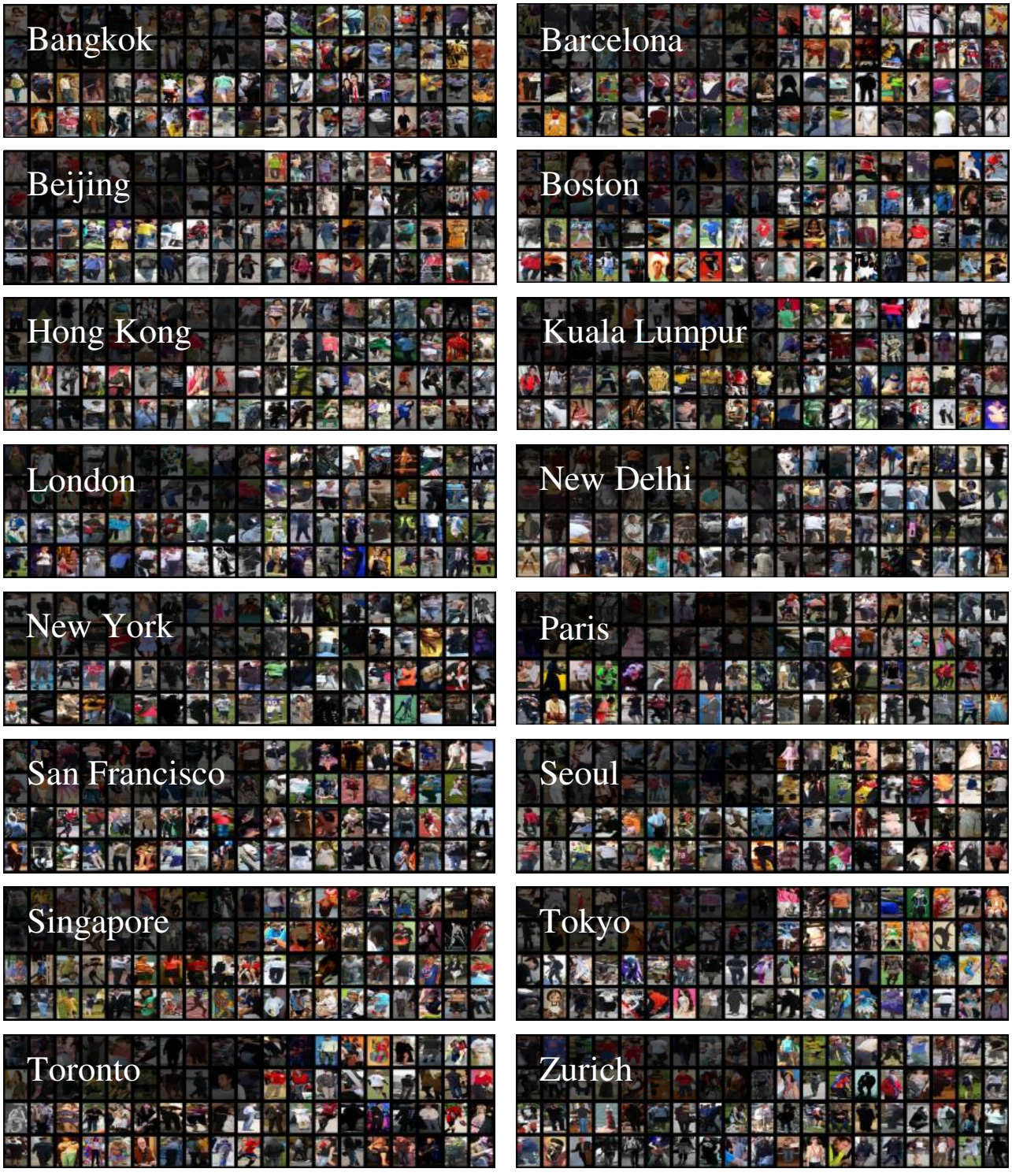


Figure 9. Example of fashion images at each city.