Transdisciplinary Engineering for Resilience: Responding to System Disruptions
L. Newnes et al. (Eds.)

© 2021 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/ATDE210088

Logo Image Retrievals Using Deep Embedding Learning

A.J.C. TRAPPEY[a,1], C.V. TRAPPEY[b] and Samuel SHIH[c,2]
[a, c] *Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu, Taiwan*
[b] *Department of Management Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan*

Abstract. A logo is s graphical emblem or mark used as an identification for a company and its products and services. Logos are legally protected as intellectual properties (IPs) if registered as trademarks (TMs). Logos™ are widely distributed online nowadays in the digital economy. Due to their wide distributions online, the constant checking of TM legal usages becomes extremely challenging in the TM registration and protection system. The fact that users can easily imitate the registered TM logo designs casts serious IP legal issue, which highlights the importance of developing an automatic logo image retrieval system. Considering the complexity of TM visual semantics, this research proposes a deep embedding learning for logo image similarity analysis using triplet-network. We propose the optimization of sampling parameters to improve the TM image retrieval performance with robust model. The research aims to reduce discrepancy between human visual interpretation. This transdisciplinary engineering research incorporates deep learning (DL) modeling and TM legal analysis for image-centric TM protection. To demonstrate the model performance, more than 10,000 images for model training and 3000 images for model testing are adopted from Logo-2K+ database. Image retrieval performance shows excellent results with recall@10 exceeding 93%.

Keywords. Image retrieval, triplet network, deep embedding learning, trademark similarity analysis, trademark infringement, transdisciplinary engineering

Introduction

With well exceeding 10 million global TM registrations per year [1], TM management and protection become challenging tasks for all IP authorities world-wide. Nowadays, although with some computer supporting systems, the world intellectual property organization (WIPO) and individual country IP offices still relies on TM examiners to check Vienna classifications to ensure no duplicated or deceptively similar TMs are registered in the IP legal system [2]. Thus, it takes a long time for the examiners to complete the registration review. TM examiners spend most of their time searching for similar TMs through the system. The examination of TM patterns tends to be subjective and can lead to inconsistencies of similarity judgement based on different subjective

[1]Corresponding author: Professor Amy Trappey, email: trappey@ie.nthu.edu.tw.
[2] Corresponding author: Samuel Shih, email: s108034571@m108.nthu.edu.tw.

views of the examiners [3][4]. Thus, computer-supported (or automatic) TM image retrieval systems become vital and indispensable.

Nowadays, machine learning (ML) methods show remarkable performances in the field of smart image retrievals, image classifications, and picturesque feature detections. However, there are still many challenges and issues while solving problems with transdisciplinary nature, e.g., the needs of TM legal/law knowledge for the ML modeling of logo^{TM} image retrievals. When the ML predictive modeling without taking TM legal characteristics into consideration, the performance of the TM similarity analysis will be largely weakened [5]. Thus, the logo^{TM} image retrieval system can fully support the accurate analysis of both TM registration/examination and TM fair use (law) judgement.

The logo^{TM} retrieval system is based on the geometric figures and visual semantics of logo designs, which can provide key information on the similarity matching of logo images, especially during the initial TM examination [6] and, also, the online logo^{TM} infringement investigation. To enhance model training efficiency and accuracy, this research adopts the pre-trained convolution neural network on the large-scale dataset and then perform fine-tuning on the small TM logo dataset. In this research, we use the VGG19 pre-trained model as the backbone from the Keras application which uses the training data from ImageNet [7]. During the model training, we construct a triplet network and improve the algorithm for selecting samples to form triplet samples. Triplet loss function is used to measure the different between the target and the predicted scalar. The model can use the loss function for back-propagating the gradient calculations. Finally, gradient descent approach is used to update the parameters of the model classifiers. In addition to the high accuracy of the logo image retrieval results, the similarity analysis can also be performed at different semantic levels.

In section 1, we discuss the literature review of image retrieval technology in recent years. Section 2 introduces the logo similarity analysis (deep learning) modeling framework and the model fine-tuning methodology. Section 3 presents the neural model verification and shows the experimental performance of the model application. Section 4 draws the conclusions.

1. Literature review

Nowadays, search engines commonly used on Internet platforms are based on keywords. In the process of keywords searching, the retrieval system uses text-based to retrieve some images or files in the database that are already labeled with specific text or some descriptions [8]. The retrieval algorithm uses text mining to analyze the similarity between keywords and tagged text. However, text-based image retrieval is not based on image features to match but uses manual labeling methods to match [9].

With the growth of ML methods in recent years, many studies focused on Content-based image retrieval (CBIR). CBIR is a method of extracting image features and transforming these visual features into specific mathematical vectors or matrices for further analysis. In the CBIR system architecture, query image is mainly used for image retrieval tasks. The system converts the query image input by the user into feature vectors that can express visual meaning and analyzes the similarity between these feature vectors and the images in the database. This study summarized the literature review of image retrieval technology in recent years into two main parts. The first part is traditional ML methods, which focuses on the low-level features of the image in the retrieval image features, including color, shape, texture or spatial layout. The other part of image retrieval technology uses DL methods to extract image features. Some studies have

shown that DL methods can extract more semantic visual concepts and save more time in algorithm performance than traditional ML methods [10][11]. In section 1.1, we introduce the application of traditional ML algorithms in the field of image retrieval. Section 1.2 introduces the development of DL in image retrieval technology.

1.1 Traditional machine learning methods

In traditional ML methods, the low-level features of images are mainly captured such as color, shape and texture. This subsection introduces the image retrieval performance of different specific mathematical models in the image retrieval literature.

Color histogram is based on statistical analysis method to measure the similarity analysis of images. Using a histogram to represent the color feature distribution in the overall image, this method can effectively represent the multi-type feature distribution but has the limitation of rotation invariance [12]. Since color information is less sensitive to changes in directionality and size, the use of color histogram analysis is relatively lacking in the local feature analysis of the image [13]. Color is an important feature information for image similarity analysis. Some studies have successfully used color distribution to extract image features for matching, which can extract image color distribution, image contrast and pixel brightness distribution [14].

Texture features can also convey important image information in image feature retrieval. Local binary patterns (LBP) are proposed by Ojala of Oulu University in Finland, which is a feature description method using regional texture changes [15]. In many practices, the advantage of the local binary pattern algorithm is that it can effectively extract image texture information. Liu et al. pointed out that LBP algorithm is weak in color feature extraction in image retrieval tasks. Therefore, a method of fusing LBP and color information feature (CIF) method is proposed to solve the problem of image classification and retrieval problem [16]. LBP was first used for texture analysis and is very effective for face recognition, expression recognition, multimedia searches, and motion analysis.

Zhou et al. purposed a collaborative index embedding to unifying index of SIFT feature and the deep convolution neural network for IR task. To show the neighborhood embedding with integrated SIFT feature and CNN feature, the index embedding algorithm will continue to update the index file of CNN and SIFT features. The indexes that are close in the SIFT space will be pulled closer in the CNN feature space. The improved CNN index output shows a significant improvement in retrieval accuracy, which is more than 10% higher than the original CNN and SIFT index [17].

1.2 Few-shot learning

Few-shot learning is a method of making predictions based on a limited number of samples. Different from standard supervised learning, few-shot learning can recognize a category that is not included in the training set [18]. In this research scope, we introduce two popular network architectures: Siamese network and Triplet network. These two simple networks can calculate similarities or distances between images in the feature space and thereby solve few-shot learning task.

Siamese Network is mainly constructed by two CNN networks. Siamese network uses paired samples to create similar and dissimilar image pairs through the binary labels and fine-tuning the model parameters through contrastive loss [19]. Some studies are focus on improving the architecture of SNN and the method of selecting training samples

[19][20]. Appalaraju et al. proposed an improvement in the method of selecting negative samples in the training model [21]. In the initial stage of model training, the easy negative sample that are easy to learn from the model should be selected, and as the training epoch increases, the hard-negative samples that are difficult to learn from the model will increase.

Triplet neural network is mainly constructed by three CNN networks. Through the Triplet loss function, images feature in the vector space strive to reduce the distance between the similar images and increase the distance between dissimilar images [22]. Wu et al. research team pointed out the importance of sampling methods in deep embedding learning and improved the original triplet loss function. In the sampling method, distance weighted sampling is proposed, which considers that each sample point in the feature space is uniformly distributed in a hypersphere space. At the same time, the distance-weighted sampling method is used to improve the effective samples for training [23]. Lan et al. research mainly compares the performance of traditional ML methods and deep learning-based methods on TM feature learning. The experimental results show that the DL method has a good performance in TM retrieval, and also shows the Triplet neural network has a better performance [24]. Min et al. constructs more complex Triplet network architecture and trained it via simultaneously optimizing the triplet loss and softmax loss during model training [25]. Veit et al. constructs Conditional Similarity Networks to learn semantic features defined by human concept during machine training [26]. Embed different semantic features into different subspaces for model similarity learning.

Compared with traditional ML methods, few-shot learning with deep embedding learning is no longer restricted to the color, shape and texture of the picture, and it can extract more semantic features in the deep neural network architecture. This research adopts the architecture of triplet network as the original logo similarity analysis DL model framework. By reorganizing data samples and improved sample selection algorithm, different visual concepts can be used to analyze the model when analysis the similarity of different logos.

2. Research methodology

Figure 1 shows the architecture diagram of the system constructed in this research. In the framework, it can be divided into three stages: pre-processing stage, training stage, testing stage and the process of model similarity analysis. The framework process at each stage will be introduced in detail in the following sections.

2.1 The pre-processing stage

In the pre-processing stage, this research includes two parts, the first part is to introduce the pre-training backbone of the neural model, and the second is how this research pre-process the TM dataset for model fine-tuning. This research imports the pre-trained CNN model from the Keras application which uses the training data from ImageNet [7]. In the CNN model architecture, three new classifiers are added, including a flatten layer, one dense layer with 512 units and output shape 512 units with L2 Norm regularizer. To improve the performance of model convergence, we add kernel initialization and regularization parameters in dense layer.

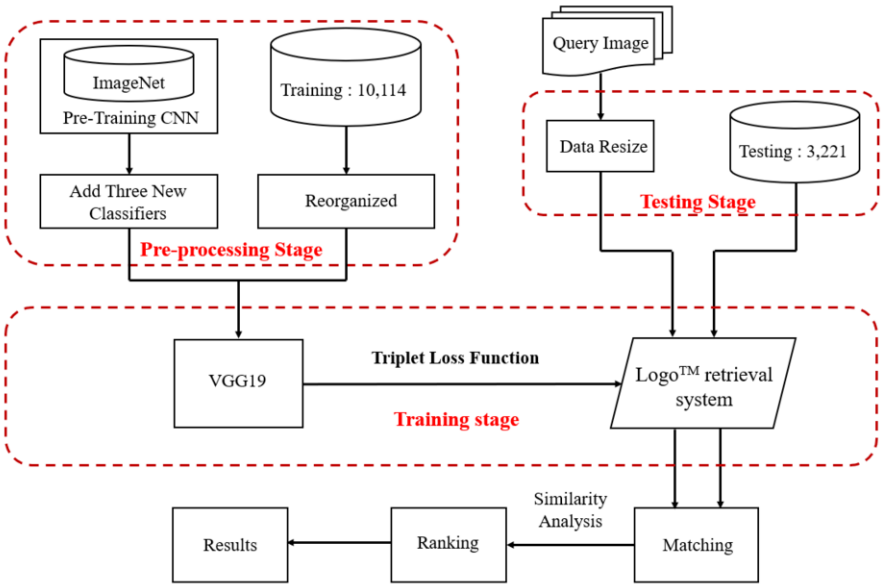


Figure 1. The framework of logo similarity prediction system.

We adopted the Logo 2K+ dataset as this research training and testing dataset. The Logo 2K+ contains 167,140 images with 10 root categories, which further sub-classified onto 2,341 company logo categories [27]. Since this research focuses on the similarity learning of geometric pattern features, we select some suitable data from Logo 2K+ as the model training and testing data set. We select TM logos with obvious similar features and clean backgrounds in the Logo 2K+ dataset as the reorganize dataset for this study. In the Logo 2K+ dataset, even in the same category, there are still some mixed company TM logos. This research follows several rules in selecting and excluding each category of TMs: (1) We choose a clean background with obvious characteristics (Figure 2). (2) Delete distortion, blur, and patterns that are too small. (3) Delete pictures with the same letters but different fonts or styles (Figure 3). (4) Select geometrically similar TM patterns in each category. The total number of TM logo reorganized in this study is 13,335 (10,114 for training, and 3,221 for testing).

Figure 2. Logos in the same category must be able to clearly show the main geometric features.

Figure 3. The TM logos belong to the same category in Logo 2K+ dataset, their geometric characteristics may be different.

2.2 Training stage of the similarity prediction model

In the training stage, this study adopts a triple network training model to learn the similarities between TM logo features. In the triple network architecture, the input channels of the model are anchor, positive, and negative three logos, in which anchor and positive are logos of the same category, and negative are the logos of different categories. The image extracts feature vectors through a pre-trained model and output the feature vectors from three new classifiers. The feature vectors of these three outputs are expressed in the feature space R^d with nonlinear f(x). Having three feature vectors, including f(a) (Anchor), f(p) (Positive) and f(n) (Negative), we can further calculate the Euclidean distance between each vector. The algorithm calculates the Euclidean distance between the two feature vectors f(a) and f(p) shown in D (f(a), f(p)) = $\|f(a) - f(p)\|_2^2$. Another distance between the two feature vectors f(a) and f(n) is shown in D (f(a), f(n)) = $\|f(a) - f(n)\|_2^2$. Further, we can use these distance values to calculate the triplet loss function (Eqn. 1). The model implements the triplet loss function to back propagation for calculating the gradients. Finally, use gradient descent to update the model parameters. Triplet loss function is to encourage the distance between feature vectors of the same category in the feature vector space to be minimized and maximizing the distance between images of different categories. Finally, the hyperparameter "Margin" is set to determine whether the positive sample and the negative sample can be distinguished in the feature vector space.

$$\mathrm{L} = \mathrm{Max}\{D\ (\mathrm{f(a)}, \mathrm{f(p)}) + \mathrm{Margin} - \ D\ (\mathrm{f(a)}, \mathrm{f(n)}), 0\} \tag{1}$$

2.3 Triplet sampling strategy for model fine-tuning

Selecting valid triplet samples will improve the performance of the training model. In this research positive sampling, we select the sample with semi-hard positive which the distance between anchor and positive is greater than the average distance of the same category. The positive sampling satisfies Eqn. 2's condition, where "p" is the number of positive samples, M(a) as the dataset with the same category as anchor "a" and f(a) and i(a) denote feature vectors.

$$D(\mathrm{f(a)}, \mathrm{f(p)}) > \frac{\sum_{i \epsilon M(a)} \|f(a) - i(a)\|_2}{p} \tag{2}$$

In negative sampling, it can be divided it into three cases including easy negative, hard negative and semi-hard negative. In easy negative situation, the model can easily distinguish between positive and negative samples, which is less helpful for model training. Hard negative sample is also difficult for the human visual to distinguish the difference between negative and anchor, and it is more difficult for model training. Selecting too many "hard negative" samples will make the model's generalization ability poor and cause the loss function difficult to converge. "Semi-hard negative" is a better interval for model training and more stable in the similarity learning. In the negative sampling strategy, first select the semi-hard negative that satisfies Eqn. 3's condition. If the first condition is not met, then we select the hard negative that satisfies Eqn. 4's condition. In Eqn. 3 and Eqn. 4, "margin" is a positive value, which defines the distance boundary between the anchor and samples.

$$D(f(a), f(p)) < D(\mathrm{f(a)}, \mathrm{f(n)}) < D(f(a), f(p)) + margin \quad (3)$$

$$D(\mathrm{f(a)}, \mathrm{f(n)}) < D(f(a), f(p)) < D(f(a), f(p)) + margin \quad (4)$$

This research sets the margin value based on the selected sample category. In the setting of the "margin", the study found that the similarity distance between the mini batch (each category) and the anchor category will be different for different categories. Therefore, the margin settings of different categories are dynamically adjusted during the negative sampling. The margin setting is the average distance of negative sample categories minus the average distance of positive samples, which show in Eqn. 5. Where "n" is the number of negative samples, "p" is the number of positive samples and N(a) is the dataset that the mini batch from anchor's category. In section 3.1, we present the performance of different sampling strategies on the triplet network model. The experimental results show that the positive sample and negative sample selection methods designed by this research perform best for TM pattern similarity learning.

$$Margin = \frac{\sum_{j \epsilon N(a)} \| f(a) - j(a) \|_2}{n} - \frac{\sum_{i \epsilon M(a)} \| f(a) - i(a) \|_2}{p} \quad (5)$$

3. Experiments

In this study, the standard Recall@K [28] metric is used to measure the logo^{TM} image retrieval system performance. Recall@K can be defined as: According to Top-K ranking results based on the similarity of a given query image, if at least one image of the same category is retrieved by the model, the recall value is 1. Otherwise, the recall value is 0.

This research constructs the neural model using Tensorflow environment [29]. The hyperparameters of the model include: batch-size (256), number of training epochs (100), input image resolution (224 × 224 pixels), triplet loss margin threshold (1.0). ADAM is the solver for optimization [30]. In the learning rate, we adopted a learning schedule. The learning rate is 0.01 in the first 15 epochs, and then the learning rate is multiplied by 0.1 every ten epochs.

3.1 Sampling strategies

In the experiment of the negative sample margin setting, there are two settings including constant margin (first three rows) and a "average margin" (last three rows) for comparison which is shown in Table 1. In positive sampling, we compare three strategies: Easy Positive, Semi Hard Positive, and Hard Positive. Easy Positive (EP) is to select the positive sample that is closest to the anchor in the feature vector space among the same category. The margin setting of the Semi-hard Positive (SP) sampling strategy depends on the average distance of the same category. The SP sampling condition is defined as Eqn. 2 which is shown in Section 2.3. In hard positive (HP) sampling strategy, the algorithm selects the farthest sample as the positive sample in the same category. In the HP strategy, the samples for model training are focused on samples that are difficult to identify, and it will increase the difficulty of model convergence. Three sampling methods can be listed as "Easy positive and Semi-hard Negative" (EPSN), "Semi-hard Positive and Semi-hard Negative" (SPSN), "Hard Positive and Semi-hard Negative" (HPSN). And we also compare with semi-hard negative sampling with constant margin

which the constant value is 0.5. Three constant margin methods can be listed as "Easy positive and Constant negative" (EPCN), "Semi-hard Positive and Constant Negative" (SPCN), "Hard Positive and Constant Negative" (HPCN).

Table 1. Model verification in positive sampling margin settings.

	Top-1		Top-5		Top-10	
Sampling Methods	mAP	Recall@1	mAP	Recall@5	mAP	Recall@10
EPCN	0.77	0.80	0.72	0.85	0.69	0.83
SPCN	0.76	0.79	0.75	0.85	0.73	0.88
HPCN	0.70	0.75	0.69	0.77	0.72	0.86
EPSN	0.79	0.81	0.74	0.85	0.73	0.89
SPSN	**0.82**	**0.84**	**0.81**	**0.89**	**0.77**	**0.93**
HPSN	0.70	0.73	0.70	0.83	0.67	0.87

From the experimental data results in Table 1, the SHSN sampling algorithm has well performance in all the retrieving results. This study proposes methods for optimizing positive and negative sample selection strategies and observes the changes in model effects in different strategies.

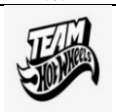
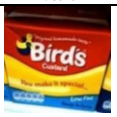
3.2 Evaluation results

Table 2 presents the results of the logo similarity predictions. The first column in Table 2 is query image, and the following five columns are the retrieval results of the model and the distance similarity presented above each TM logo. We can see that when the similarity distance calculated by the model is less than 0.5, it can be said that it is directly similar to the overall appearance of the query image. For the case where the similarity distance is greater than 0.5, it can be known that some features are indirectly similar to the query image. Some factors that can cause bias in the logo^{TM} image retrieval system results are: color changes, background noise, and image clarity.

4. Conclusion

Based on the triplet network architecture, we built a deep neural network model focusing on the learning of multi-concept similarity of TM patterns. The contribution of the logo^{TM} image retrieval system in this research is to optimize the network parameters and improve the sampling strategy based on the similarity learning of TM features in the existing triplet network architecture. In the verification, the parameter optimization, based on SPSN, yields the best performance in learning the similarity of TM features. In addition to improving the retrieval accuracy of the model, the distance analyzed in the logo^{TM} image retrieval system can also be used to determine whether there is visual semantic similarity. The model trained and tested in this research provides the computer supported inspection for accurate and efficient TM registration and protection processes.

Table 2. The logo™ image retrieval system experiment results.

0.0	0.5	0.591	0.596	0.62	0.63
0.0	0.64	0.65	0.69	0.71	0.73
					
0.0	0.39	0.51	0.61	0.67	0.67
0.0	0.11	0.39	0.68	0.71	0.73
0.0	0.38	0.50	0.55	0.58	0.63
					

Acknowledgement

The research is partially supported by the research grants of Ministry of Science and Technology, Taiwan (Grant numbers: MOST-108-443 2221-E-007-075-MY3 and MOST-108-2410-H-009-025-MY2).

References

[1] WIPO, 2019, *World Intellectual Property Indicators 2019: World Intellectual Property Organization.* [Online]. Available: https://www.wipo.int/publications/en/details.jsp?id=4464/

[2] J. Schietse, J.P. Eakins and R.C. Veltkamp, Practice and challenges in trademark image retrieval, *Proceedings of the 6th ACM international conference on Image and video retrieval*, 2007, pp. 518-524.

[3] M.D. Frakes and M.F. Wasserman, Are There as Many Trademark Offices as Trademark Examiners? 2020, *Duke Law Journal,* 69, 1807-1853.

[4] J.T. McCarthy, *McCarthy on Trademarks and Unfair Competition*, 5th ed, Thomson West, Deerfield, 2020.

[5] S. McJohn and I. McJohn, Fair Use and Machine Learning, *Northeastern University Law Review, Forthcoming*, 2019, https://ssrn.com/abstract=3406283, Accessed: July, 3 2021.

[6] Kinney and Lange, P. A., *Intellectual Property Law for Business Lawyers,* Thomson West, Eagan, 2020.

[7] C. François, Keras, 2019, Accessed: 30.05.2019. [Online]. *Avaiable: https://keras.io/*

[8] S. Yang, L. Li, S. Wang, W. Zhang, Q. Huang, and Q. Tian, Skeletonnet: A hybrid network with a skeleton-embedding process for multi-view image representation learning, *IEEE Transactions on Multimedia,* 2019, 21(11), pp. 2916-2929.

[9] A. Amelio, A new axiomatic methodology for the image similarity, *Applied Soft Computing*, 2019, Vol. 81, 105474.

[10] U. Markowska-Kaczmar and H. Kwaśnicka, Deep Learning—A New Era in Bridging the Semantic Gap, In: H. Kwaśnicka, L.C. Jain (eds.) *Bridging the Semantic Gap in Image and Video Analysis*, Springer Nature Switzerland, 2018, pp. 123-159.

[11] G. Qi, H. Wang, M. Haner, C. Weng, S. Chen and Z. Zhu, Convolutional neural network based detection and judgement of environmental obstacle in vehicle operation, *CAAI Transactions on Intelligence Technology,* 2019, 4(2), pp. 80-91.

[12] K. Roy and J. Mukherjee, Image similarity measure using color histogram, color coherence vector, and sobel method, *International Journal of Science and Research (IJSR),* 2013, 2(1), pp. 538-543.

[13] R. Chakravarti and X. Meng, A study of color histogram based image retrieval, *2009 Sixth International Conference on Information Technology: New Generations,* 2009, pp. 1323-1328.

[14] J.-M. Guo, H. Prasetyo and J.-H. Chen, Content-based image retrieval using error diffusion block truncation coding features, *IEEE Transactions on Circuits and Systems for Video Technology,* 2014, 25(3), pp. 466-481.

[15] T. Ojala, M. Pietikainen and T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, *IEEE Transactions on pattern analysis and machine intelligence,* 2002, Vol. 24(7), pp. 971-987.

[16] P. Liu, J.-M. Guo, K. Chamnongthai and H. Prasetyo, Fusion of color histogram and LBP-based features for texture image retrieval and classification, *Information Sciences,* 2017, Vol. 390, pp. 95-111.

[17] W. Zhou, H. Li, J. Sun and Q. Tian, Collaborative index embedding for image retrieval, *IEEE Transactions on pattern analysis and machine intelligence,* 2017, Vol. 40(5), pp. 1154-1166.

[18] L. Fei-Fei, R. Fergus and P. Perona, One-shot learning of object categories, *IEEE Transactions on pattern analysis and machine intelligence*, 2006, Vol. 28(4), pp. 594-611.

[19] S. Zagoruyko and N. Komodakis, Learning to compare image patches via convolutional neural networks, *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015, DOI: 10.1109/CVPR.2015.7299064.

[20] C.V. Trappey, A.J. Trappey and S.C.-C. Lin, Intelligent trademark similarity analysis of image, spelling, and phonetic features using machine learning methodologies, *Advanced Engineering Informatics*, 2020, Vol. 45, 101120.

[21] S. Appalaraju and V. Chaoji, Image similarity using deep CNN and curriculum learning, *arXiv preprint arXiv:1709.08761*, 2017.

[22] F. Schroff, D. Kalenichenko and J. Philbin, Facenet: A unified embedding for face recognition and clustering, *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015, DOI: 10.1109/CVPR.2015.7298682.

[23] C.-Y. Wu, R. Manmatha, A.J. Smola and P. Krahenbuhl, Sampling matters in deep embedding learning, *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, 2017, DOI: 10.1109/ICCV.2017.309.

[24] T. Lan, X. Feng, L. Li and Z. Xia, Similar trademark image retrieval based on convolutional neural network and constraint theory, *2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 2018, DOI: 10.1109/IPTA.2018.8608162.

[25] W. Min, S. Mei, Z. Li and S. Jiang, A Two-Stage Triplet Network Training Framework for Image Retrieval, *IEEE Transactions on Multimedia*, 2020, Vol. 22, Issue 12, Dec. 2020, pp. 3128 - 3138.

[26] A. Veit, S. Belongie and T. Karaletsos, Conditional similarity networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, DOI: 10.1109/CVPR.2017.193.

[27] J. Wang, W. Min, S. Hou, S. Ma, Y. Zheng, H. Wang and S. Jiang, Logo-2K+: A Large-Scale Logo Dataset for Scalable Logo Classification, *AAAI*, 2020, https://doi.org/10.1609/aaai.v34i04.6085.

[28] H. Oh Song, Y. Xiang, S. Jegelka and S., Savarese, Deep metric learning via lifted structured feature embedding, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4004-4012.

[29] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, . . . M. Isard, Tensorflow: A system for large-scale machine learning, *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265-283.

[30] D.P. Kingma and J. Ba, Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980*, 2014.