March - April 2020
ISSN: 0193-4120 Page No. 4503 - 4510

A Study on Image Retrieval System for Clothing Materials Using Convolutional VAE

Yeonghun Lee[1], Jonghyung Sung[2], Hyunghwa Ko[3], Kyounghak Lee[*4]
[1]Research Scholar, [1]Department of Electronics and Communications Engineering, Kwangwoon University, Seoul, 01897, Korea.
[2]Research Scholar, Department of Research and Development, JongDal Lab Co. Ltd., Seoul, 01897, Korea.
[3]Professor, Department of Electronics and Communications Engineering, Kwangwoon University, Seoul, 01897, Korea.
[*4]Associate Professor, IACF, Kwangwoon University, Seoul, 01897, Korea.
yeonghun237@kw.ac.kr[1], cresper@jongdali.com[2], hhkoh@kw.ac.kr[3], goldbug@kw.ac.kr[*4]

Article Info
Volume 83
Page Number: 4503 - 4510
Publication Issue:
March - April 2020

Abstract
Recent Image Processing with Deep Learning has been permeated in our daily life. In accordance with development of Deep Learning, we attempt to apply it to Clothing Materials Retrieval.

Proposed model consists with YOLO detecting clothing materials and Convolutional VAE characterizing images in database. Since the vectorized image feature is high-dimensional, PCA is applied to reduce features of image and time for retrieval. Furthermore, KNN is utilized to search for k-similar images in reduced vector space. To test this system, we collected dataset by Web Crawling and its result shows 10-near images for each arbitrary test image.

YOLO is efficient model to produce an image with exactly necessary region only. As an Object Detector, it achieves to generate cropped images appropriately from poor dataset. VAE needs huge dataset in training stage and Encoder of VAE amply works as an image discriminator. Even though we reduced dimension of vectorized features from 512 to 73 or lower dimensions by PCA, it does not have terrific loss. It rather brings reduction of an image search response. KNN is suitable distance measure method when sort out similar images in comparatively low dimension. Importance of our study is on implementation of clothing materials retrieval system by combining Object Detection as region of interest, unsupervised Deep Learning, PCA and KNN algorithms. Proposed model favorably retrieves similar clothing materials based on color and texture without any labeling process in real-time.

Our image retrieval system for clothing materials results 0.1 second for 10-similar image retrieval from 170,681 images on 73-dimensional latent space.

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 26 March 2020*

Keywords: *Image Processing, Deep Learning, Clothing Materials, Image Retrieval, Unsupervised*

1. Introduction

Image Processing is powerful tool for smart factory and object recognition for autonomous car. Along with development of Deep Learning, this technology has conducted fundamental role. One of applications using image processing is image retrieval such as Google Images, Clarifai Visual Search and so on. Those systems are generalized and cannot cover all kinds of object. Hence, it is required specific dataset. Generally, datasets are collected from the internet, but images are so unrefined that it includes chaotic background with

March - April 2020
ISSN: 0193-4120 Page No. 4503 - 4510

diversity of angles, objects, colors and several conditions. It leads to extract even feature of useless area. Therefore, refining data and extracting feature of object region are demanded to search a relevant image.

To refine image dataset for deep learning training depending on requirement, irrelevant part should be removed. Segmentation such as DeepLab could be one of methods to acquire only target we need by getting rid of background. But it needs high cost in computation and manpower. Alternatively, Object Detection making boundary box can be used to make cropped image data, although it still includes background. Object Detection models can be categorized into two methods, two-stage and one-stage. Two-stage method such as Faster R-CNN is comparatively slower and more inaccurate than one-stage method such as SDD. Among some popular one-stage Object Detection model, YOLOv3 is fastest and accurate. Hence, it would be most suitable for real-time Image Retrieval[1,2,3,4].

Since image searching system has been conducted with keyword on internet, automatic image annotation has been getting studied. Another approach for Image Retrieval is based on contents of image matching a query image. It is commonly called content-based image retrieval (CBIR) and overcome limitation of keyword-based approach. The compressed content feature is placed in database and offer similar images by selecting nearest feature vector out. However, the representation of content is usually high dimensional due to reliable discrimination. To reduce computational cost by dimension reduction, Principle Component Analysis (PCA) has been mainly employed. Some study on PCA for massive dataset describes method that representation with few local principal components can be substituted with small accuracy decrement[5,6,7,8].

Recent years have been highly active in Convolutional Neural Network (CNN) due to its effectiveness for image. The convolution has been greatly efficient as filter for decades in computer vision field to extract edge, shape and texture, which is same principal of CNN filter. Further, CNN extracts high-level features so that covers semantic discrepancy between low-level image by machine and high-level by human. Thus, enormous works have dealt with Image Retrieval using CNN, and now it became a trend[9].

Our goal is an unsupervised model in searching stage to retrieve similar image and use Variational Auto Encoder (VAE). As shown above that CNN is efficient for image, we adopt Convolutional Variational Auto Encoder (CVAE). Investigation of Auto Encoder learning representation without labeling which has equal classification accuracy as much as CNN, is embraced in analysis study. Encoder of VAE is trained to compress content as a feature vector in latent space and Decoder restores input image from the feature vector. It means that Encoder part carries out role of embedding feature into most worthwhile values in latent space. On the other hand, Auto Encoder cannot be sufficiently trained with low dimensional latent space. Consequently, the vectorized feature has hundreds of parameters, which is high-dimensional. That is why PCA follows CBIR for dimension reduction[10].

Moreover, distance measure method of Euclidean Distance would be meaningless in high dimension. Cosine similarity is alternative. But KNN[11] would be best in reduced dimension.

2. Methods

Proposed model consists of YOLO, CVAE, PCA and KNN as distance measure.

• Database Generation

We collected 29,094 button images including plural objects and refine the images into 170,681 images by YOLOv3. CVAE weights are trained by those images, and only Encoder is used for image

March - April 2020
ISSN: 0193-4120 Page No. 4503 - 4510

vectorization. PCA is utilized for dimensional reduction of database and test images. In test step, KNN offers k-similar images for an input image.

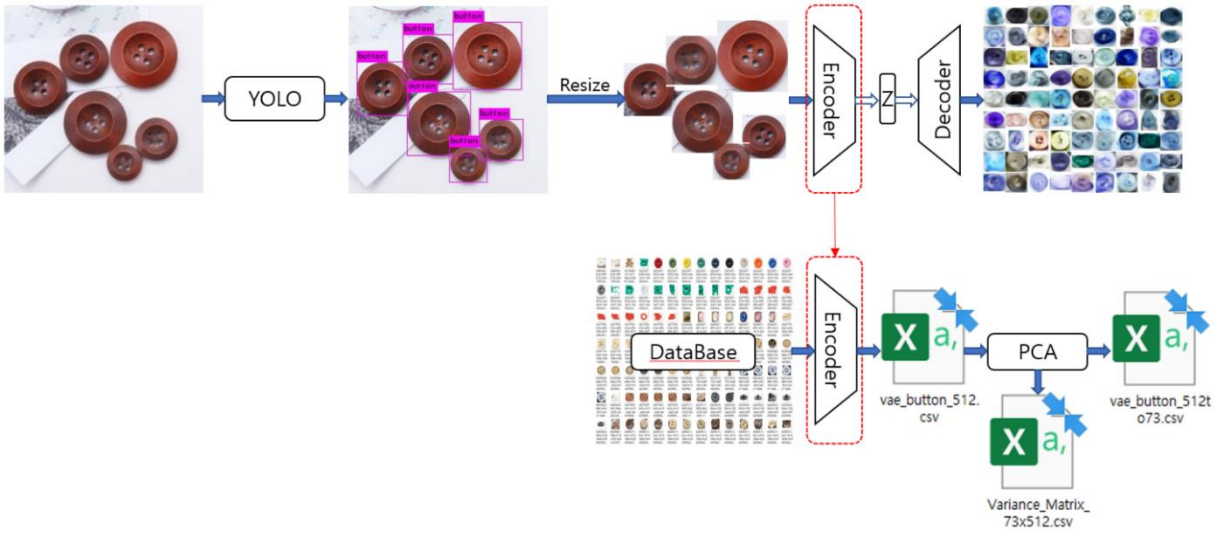

Figure 1. Data Generation Structure: 1.Object Detection 2.Vectorization 3.Dimension Reduction

Figure 1 depict process to make a training image set and vectorized features in latent space. Above input image of YOLO is one of examples and describes object detection step. As the result boundaries, objects are divided to distinct object and it becomes our database as well. Then, CVAE is trained with refined training images and Encoder part is used to create 512-dimensional features. In last step, PCA functions to reduce computational cost by decreasing vector parameters.

• Data Refinement

The arbitrary collected image is complex and even include multiple object. Thus, dataset is required to be refined as requirement. Since we focus on clothing materials, we should select images and take object part only. Object Detection is one of way to extract object. Accordingly, we proposed to combine YOLO with CBIR. We pursuit real-time searching system and regard YOLO as suited our purpose. Under the YOLO official site, this model is fastest one with near state-of-the-art accuracy. Performance comparison of Object Detection(OD) models is in table 1.

Table 1. OD Model Comparison[12,13,14]

Model	mAP	FLOPS	FPS
R-FCN	51.9	-	12
FPN FRCN	59.1	-	6
Retinanet-10-500	50.9	-	14
Retinanet-101-500	53.1	-	11
SSD-300	41.2	-	46
YOLOv2-608	48.1	62.94	40
YOLOv3-320	51.5	38.97	45

Grid system of YOLO is depicted in Figure 2. Each grid cell has bounding box, confidence for the box and class probability. It allows predicting object without region proposal stage in real-time.

Its accuracy relies on amount of dataset, which demands a lot of time to make an object boundary box. We collected clothing material images of other classes such as zipper, thread, string, race, wappen (or logo) and so on though. For practicality test, we used only button image and small labeled dataset.

We separated 29,094 images into 300 labeled image with boundary information for training and 28,794 images for YOLO input. As shown in

March - April 2020
ISSN: 0193-4120 Page No. 4503 - 4510

Figure 3, we got 170,681 images from 28,794 images. It would be also applied to other classes.

abovementioned reasons in introduction part, we apply Convolutional network to Auto Encoder.

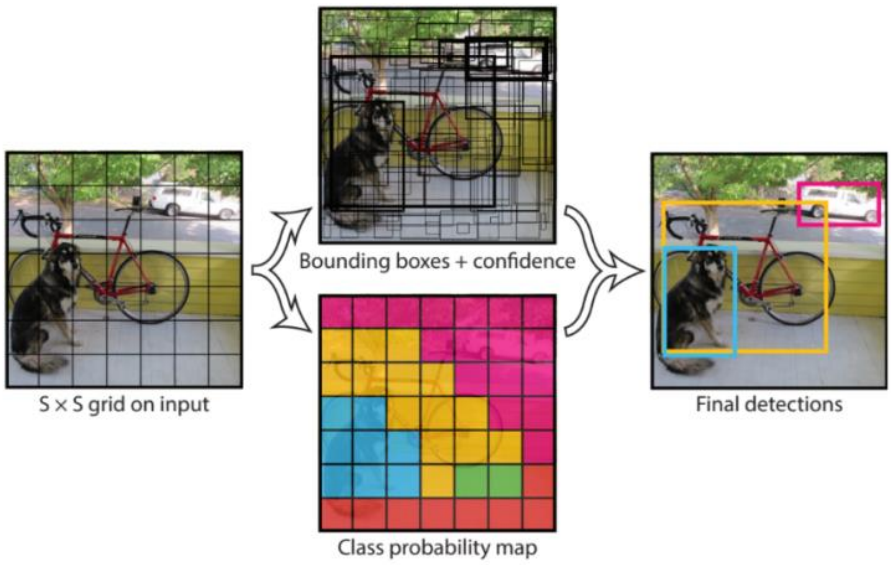

Figure 2. YOLO model: grid system[15]

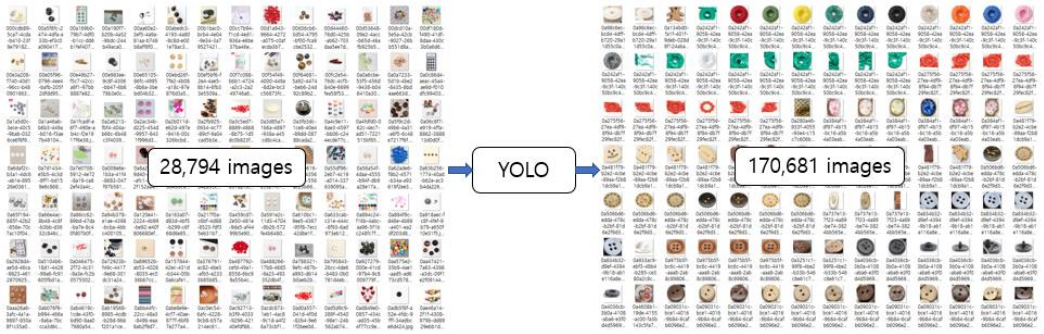

Figure 3. Extracted clothing materials by YOLOv3

• Vectorization

Content-based Image Retrieval matches original image to query image. The query images include compact features only, which represent original image. Since it manifests vector shape with dimensionality in latent space, we named the encoding stage as Vectorization.

Image data is so immense that we can deal database with unsupervised deep learning model for convenience. A representative model is Auto Encoder (AE) that encode data to compressed representation and decode it to original data. Encoder of AE is powerful method to compress input image into most relevant features. As

Moreover, Variational AE can train probability model of latent space by regularization. It leads to have better properties in latent space. In addition, we expect that CVAE filter irregular data out in searching similarity by class.

We built CVAE model to compress 512-dimensional latent value. Figure 4 depict Encoder based on general Convolution Auto Encoder. Convolutional layer is 10 layers, repeating stride size 1 and 2. Input size is 64x64 size considering heavy parameters of model. Decoder is symmetrical structure with encoder. To implement Variational AE, we added sampling stage converting input image into 512-dimensional latent

March - April 2020
ISSN: 0193-4120 Page No. 4503 - 4510

space with two parameters, which indicate a mean and variance of Gaussian distribution. Then, the sampled arbitrary point in latent space is passed to Decode and its results are match Encoder input. In result Loss function are Reconstruction Loss, KL Divergence of Latent Distribution Loss. CVAE Loss consist with the two loss. CVAE model is based on existing work. With trained weight of Encoder, 170,681 images are vectorized and saved in csv file as refined database[16].

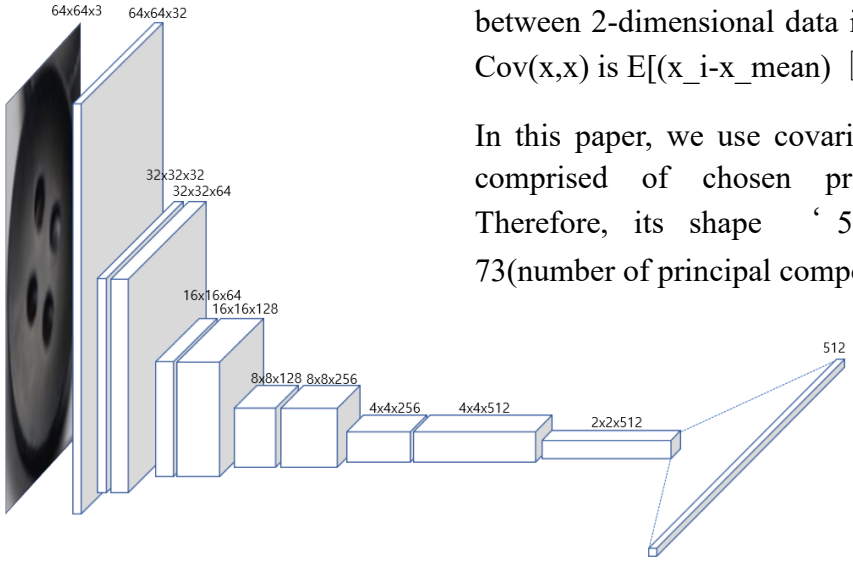

Figure 4. Convolutional VAE Encoder part structure

• Dimension Reduction

PCA is a relatively popular method for people. Principlal component means biggest variance direction vector. That is, principal component explains possible variability and represents common features of data. Therefore, it has been utilized in wide applications such as Eigen Face, Denoise, Dimension Reduction. Since 512-dimensional vector is fairly on high space, we bring PCA algorithm to reduce computational complexity problem. This algorithm workflow as follows: 1. Data Normalization. 2. Principal components Analysis in order of importance. 3. Selection of optimal principal components number. Principal component can be obtained by calculating covariance matrix. Covariance is variance to describe interrelated distribution shape between variates. Accordingly, dimension can be handled as variate and projected by covariance matrix. As equation maximizing variance, covariance matrix defined by eigen decomposition. For more information, PCA material will give you great understanding[17].

While abovementioned covariance matrix means n x n matrix, which describe correlation between dimensions. For instance, covariance matrix C=[[Cov(x,x),Cov(x,y)],[Cov(x,y),Cov(y,y)]] between 2-dimensional data is 2x2 Matrix, where Cov(x,x) is E[(x_i-x_mean)〖(x_j-x_mean)〗^T].

In this paper, we use covariance matrix term as comprised of chosen principal component. Therefore, its shape ‘512(latent vector) x 73(number of principal component)’.

For more additional grasp, Figure 5 shows our dataset projected into 2-dimensional space.

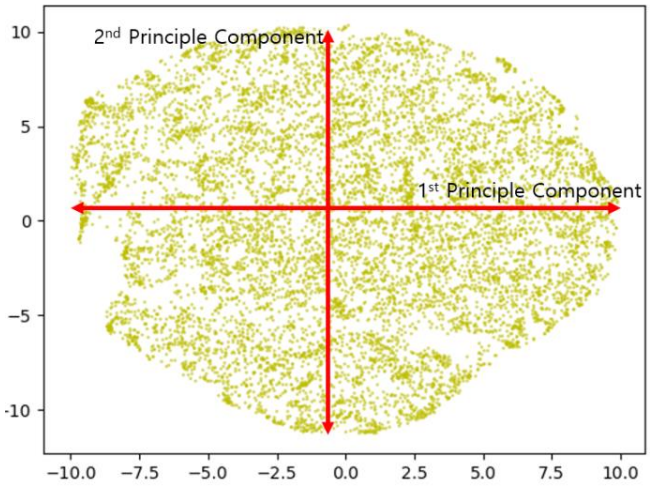

Figure 5. PCA: distribution of button dataset in 2-demensional space

March - April 2020
ISSN: 0193-4120 Page No. 4503 - 4510

As depicted above, first principle component has largest variance. It is regarded as a base vector, which is orthogonal against each principle component. At same time computing covariance matrix of 170,681, database contents are compressed to 73-dimensional vector in csv file. Dimension convertor, covariance matrix 73x512, is saved and used for test image.

• Similarity Retrieval

To retrieve similar image with input, distance measure method is required. Generally, Euclidean Distance is used to measure span between targets but, it is vailed in high-dimensional space. Thus, Cosine similarity is an alternative tool. However, we expect that KNN algorithm based on distance measure sweetly works in reduced dimension. There is no general best distance metric. Comparison review may give you clearance about distance metrics[18].

3. Results

Proposed model trained in this configuration: Window 10, Intel core i5-6600 CPU and Nvidia GeForce 1050Ti GPU. We focus on test how image retrieval performs. CVAE is trained in batch size 64, epoch 100 with 170,681 button images for 10 hours. Figure 6 shows 5 similar images for each 8 test images.

In figure 6, Left result relatively retrieve quite similar image in terms of color and shape. On the other hand, pictures on right side show poor results. When it comes to search complex shape, limit is obvious. Flower shape is estimated star-shaped one and annual-ring-shaped image is close to various kind of button in our database. The last two image on right figure seem to be embedded into wrong place in latent space. Especially last text printed button shows lack of data. 73-dimensional vector can explain 95% of accumulated variance in our latent space, which means 73 parameters includes 95% features of database.

Since this dataset is not labeled, accuracy metric is not used for assessment. Thus, we tested in term of qualitative evaluation, but execution time is included due to importance in Retrieval. Next Figure 7 shows variance rate change against 512-dimensional original vector, according to dimension reduction by PCA.

In figure 7 shows a decline from 100% in 512-dimensional feature space to 30%. Then as a dimension selection step, we refer to Loss Ratio by dimension reduction and choose 73-dimensional space, which has largest drop rate in dimension with approximately 5% loss of variance.

Figure 6. 73-dimensional space data base result: Good cases(left), Bad cases(right)

March - April 2020
ISSN: 0193-4120 Page No. 4503 - 4510

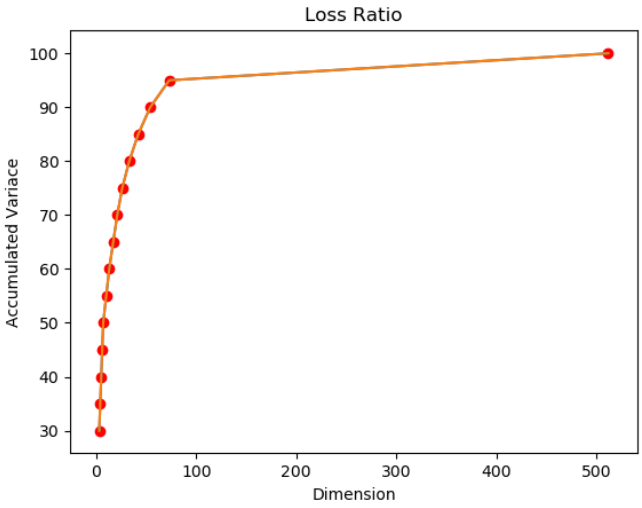

Figure 7. Loss ratio by Dimension Reduction

When image retrieval of 10 similarity among 170,681 images in 73-dimensional space, it takes 0.0897 seconds, whereas 512-dimensional retrieval takes 0.2713 seconds. In term of speed, it outperforms 2-3 times. More details of speed are shown in Table 2.

Table 2. The result of operation speed

Stage / n-D Latent Space	512-D(sec)	73-D(sec)
KNN Classifier	54.6841	**5.5611**
Encoding	3.0264	3.0278
Dimension Reduction	0	0.0409
Image Retrieval	0.2520	**0.0917**

Even database reading time of KNN is 10-times faster, though it is not big problem if server is always running. In other word, it is not surprising improvement. But remind that the more gathering data, the more time will be spent. Further we collect only 28,794 images and its quality extremely varied, which mean it could be improved. In addition, detection time of Yolo takes average 0.03 seconds in image including 3-4 objects.

4. Conclusion

Our idea is to combine YOLO for image refinement and CVAE for Embedding features on latent space. YOLO generates proper image data for CVAE training. Present work dealt with only one class, button, without classification but other classes such as thread, zipper, string, lace, band and so on will be included in our future work. Trained Encoder of CVAE properly carried a roll out embedding feature on latent space and create 900MB size database with 170,681 images. PCA allow database size to reduce to 300MB and be 2-3 times faster with closely same accuracy. In this paper, proposed model showed worthwhile potential as Image Retrieval system. As a future work, we will extend this work to other clothing materials and improve it by massive dataset from clothing industries and data refinement process.

5. Acknowledgment

This study was supported by the Seoul R&BD Program (CY190026) funded by SBA.

References

[1] Liang-Chieh C, George P,Jasonas K, Kevin M, Alan LY. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRF. Computer Vision and Patter Recognition (CVPR). 2016 Jun; Conference paper. arXiv:1680.00915.

[2] Shaoquing R, Kaiming H, Ross G, Jian S. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. Computer Vision and Patter Recognition (CVPR). 2015 Jun; Conference paper. arXiv:1506.01497

[3] Wei L, Dragomir A, Dumitru E, Christian S, Scott R, Cheng-Yang F, et al. SSD: Single Shot MultiBox Detector. European Conference on Computer Vision (ECCV) and CVPR. 2016 Dec; ECCV2016:21-37. DOI: 10.1007/978-3-319-46448-0_2.

[4] Joseph R, Ali F. YOLOv3: An Incremental Improvement. Computer Vision and Patter Recognition (CVPR). 2018 Apr; Conference paper. arXiv:1804.02767.

[5] J. Jeon, V. Lavrenko, R. Manmatha. Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval.

2003 Jul; SIGIR '03:119-126. DOI: 10.1145/860435.860459.

[6] Xiao K, Jiawei Z, Yuzhen N. End-to-End Automatic Image Annotation Based on Deep CNN and Multi-Label Data Augmentation. IEEE Transactions on Multimedia. 2019 Jan; 21(8):2093-2106. DOI: 10.1109/TMM.2019.2895511.

[7] Avinash NB, B. B. Meshram. Content Based Image Indexing and Retrieval. International Journal of Graphics & Image Processing (IJGIP) and CVPR. 2014 Jan; 3(4):235-246.

[8] Yongming Q, George O, Nagiza S, Al G. Principal Component Analysis for Dimension Reduction in Massive Distributed Data Sets. Second SIAM International Conference on Data Mining. 2002 April; Conference paper. Available from: https://www.researchgate.net/publication/232063041.

[9] Ruigang F, Biao L, Yinghui G, Ping W. Content-Based Image Retrieval Based on CNN and SVM. IEEE International Conference on Computer and Communications (ICCC). 2016 Oct; Conference paper. DOI: 10.1109/CompComm.2016.7924779

[10] Gabriel BC, Leonardo SFR, Moacir AP. Unsupervised representation learning using convolutional and stacked auto-encoders: a domain and cross-domain feature space analysis. Computer Vision and Pattern Recognition. 2018 Nov; Conference paper. arXiv:1811.00473

[11] Sahibsingh AD. The Distance-Weighted k-Nearest-Neighbor Rule. IEEE Transactions on Systems, Man, and Cybernetics. 1976 Apr; SMC6(4):325-327. DOI: 10.1109/TSMC.1976.5408784

[12] Jifeng D, Yi L, Kaiming H, Jian s. R-FCN: Object Detection via Region-based Fully Convolutional Networks. Computer Vision and Pattern Recognition (CVPR). 2016 May; Conference Paper. arXiv:1605.06409

[13] Tsung-Yi L, Piotr D, Ross G, Kaiming B, Bharath H, Serge B. Computer Vision and Pattern Recognition (CVPR). 2016 Dec; Conference Paper. arXiv:1612.03144

[14] Tsung-Yi L, Priya G, Ross G, Kaiming H, Piotr D. Focal Loss for Dense Object Detection Computer Vision and Pattern Recognition (CVPR). 2017 Aug; Conference Paper. arXiv:1708.02002

[15] Joseph R, Santosh D, Ross G, Ali F. You Only Look Once: Unified, Real-Time Object Detection. Computer Vision and Patter Recognition (CVPR). 2016 Dec; Conference paper. arXiv:1612.08242

[16] Xianxu H, Linlin S, Ke S, Guoping Q. Deep Feature Consistent Variational Autoencoder. Computer Vision and Patter Recognition (CVPR). 2016 Oct; Conference paper. arXiv:1610.00291

[17] Jonathon S. A Tutorial on Principal Component Analysis. Carnegie Mellon University. 2005 Dec; Educational material. Available: https://www.cs.cmu.edu/~elaw/papers/pca.pdf

[18] V. B. Surya P, Haneen AAA, Ahmad BAH, Omar L, Ahmad ST, Mahmoud BA, et al. Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier -- A Review. 2019 Sep; arXiv:1708.04321