An Effective Approach to Image Embeddings for E-Commerce

Volodymyr Kubytskyi [1,2] and Taras Panchenko [1]

[1] *Taras Shevchenko National University of Kyiv, Akademika Glushkova ave., 4d, Kyiv, Ukraine*
[2] *LUN.UA, Kyiv, Ukraine*

Abstract
There are lots of images in the e-commerce world. They need to be analyzed, classified, stored efficiently, compared, described in the text, searched through, and so on. The task of image de-duplication or searching near similar images is important and challenging. One of the efficient approaches to these tasks is to have an image descriptor, which helps to identify distinctive features of pictures and to organize them. The model of a such descriptor is proposed in this work. Here we describe its structure and the successful experience of its application to real-life tasks in LUN.UA. Corresponding measurements of effectiveness in comparison with other approaches are also provided. The F_1 score appeared to be higher for the proposed model. The estimation and future work are also fixed.

Keywords [1]
eCommerce, image embedding, image representation vector, convolutional neural network, CNN, CNN layer combination, image classification, image descriptor, image deduplication, near duplicate image search, de-duplication problem

1. Introduction

The e-commerce sphere consists of many topics, including internet marketing, automated data gathering, and others. It deals a lot with different kinds of multimedia materials: images and videos. Considering the huge amount of such kind of data, there are many challenges to dealing with these data in an efficient way: to store, to process, to find, and so on. The modern world produces hundreds of millions of images every day. There is a general question about the possibility of "comparing" images with each other. For example, to optimize storage because, on hundreds of terabytes of data, optimization saves thousands of dollars. Also, qualitative image embedding helps to solve the tasks of image "stylistics" recognition, scanned copies of documents classification, visual navigation, identification of diseases from X-ray or MRI images, and even three-dimensional reconstruction from a set of two-dimensional images. In e-commerce we have such situations everywhere:

- e-commerce platforms wish to deny the copies of the same goods (or to group them), which could be identified by comparing the images particularly, – or, at least, to aggregate such "duplicates";

- it's a good idea to have text descriptions for images (image-to-text, image description or annotation, or title generation task) to compare automatically with the given description, – or just to have it;

- de-duplication of images, goods, advertisements, etc. – to make the content more systemized, to provide the end user with a better experience, for the platform to look better and more solid, and also, on the other hand, to organize sellers, to avoid cheating and prohibited behavior on the platform;

- search by image functionality could be of real interest for e-commerce platforms, to give a user the functionality of finding similar goods.

For example, LUN.UA, a leading e-commerce platform in Ukrainian real estate, faced the same problems of image and advertisement duplication on their platform, and it was a crucial point and one of the key problems – to organize all real estate objects efficiently – to coin this research. So, we use

this case as an exhaustive example for our research and development. Image de-duplication or near duplicate search [1] has been an important issue for many tasks and a much broader problem indeed, and there is much research on this topic for different applications, for example:

- web search [2-7], images search [2,3] and video search [4-6], and even web documents as a whole [7];
- consumer and personal photo management [8,9];
- images clustering [10,11];
- semantic indexing [12] and deep semantic features analysis [13];
- real-time image protection and analysis [2,14] and large-scale high-load and fast analysis [15,16];
- IoT applications, for example for visual sensors [17];
- in biology [18], medicine [19], and agriculture (cropping);
- for plagiarism detection [20];
- and even for spam detection [21,22].

Google and other services provide search by image, the functionality powered by near duplicate images search behind the scene. In our research, we are aimed to solve the near duplicate image search task. We have prepared the appropriate dataset and have built an adequate model as a solution. Also, we measured the efficiency of the model proposed. As a result, the descriptive image vector is built by combining different levels of layers of a constructed and modified Convolutional Neural Network (CNN), a kind of artificial neural network. In this work, we describe the structure of the solution proposed, the results obtained, and its benefits based on the application in LUN.UA for image classification problems. Also, we draft the next research items on the topic.

2. The image embedding model proposed

To solve the issues mentioned in the Introduction section, the universal image embedding system was developed. And its application helped to solve the challenges we faced in LUN.UA. There are existing approaches, and some of them are described in accessible sources, while some others are with "closed sources". They are based on different ideas and heuristics:

- sub-image retrieval [23,24];
- local-based binary representation [25];
- keyframe identification with interest point matching and pattern learning [26];
- keypoint-based with scale-rotation invariant pattern entropy analysis [27];
- geometric invariant features [28];
- color histogram, local complexity based on entropy [29], which is fast enough, as authors claim;
- min-hash and TF-IDF weighting [30], other signatures [31];
- affinity propagation [32];
- CNN-based methods and ideas [33-35]: global and local features matching, and intermediate layers aggregation;
- colour histograms and locality sensitive hashing [36], SIFT [36], approximate set intersections between documents computing [36] and coined new datasets for state-of-the-art methods development and benchmarks for progress tracking [9,18,37-40].

There is also a common approach for such a class of tasks named "embeddings" [41-46]. This means building a description numeric vector for each image, which is distinctive enough to catch the specifics of each particular image. We can find similar models ("embeddings") for texts, audio, and other kinds of media. The developed image embedding system with a decision block consists of 3 nodes:
1. Image feature extractor – embedding builder.
2. Distributed embeddings storage.
3. Decision-making unit.
Let's describe it in more detail.

2.1. Image feature extractor – embedding builder

The extractor is formed by inheriting a pre-trained convolutional neural network for image classification – its architecture and weights are used. For example, let's take a ResNet50 (residual neural

network) [47] architecture having multiple convolutional blocks and downsampling blocks. We take a pre-trained model, trained to recognize 1000 ImageNet [48] categories.

The fully connected output layer responsible for the image class is removed from the existing network. The critical point of the image feature extractor is the union of N-intermediate layers of a convolutional neural network into a single resulting vector. This step is essential to obtain a qualitatively new level of image description. Because, on different layers of the CNN, we have highlighted different feature types – from the most abstract at the beginning to more specific ones at the end. So, on the initial layers of the convolutional neural network, basic shapes are selected (point, line, circle), and towards the end, the layers can choose complex objects and attributes (teapot, sofa, iron).

Concatenating convolutional layers with different receptive image fields into a single normalized vector is one of the ways that make it possible to form a vector of characteristics sufficient for comparisons. The resulting extractor is resistant to linear image transformations, brightness, contrast, and rotation changes by a given angle, and is also insensitive to image noise and watermarks.

An image is given at the input of the characteristics extractor. The output is a one-dimensional vector of real numbers describing the characteristics of the picture – image embedding.

The proposed extractor can universally describe the critical characteristics of images. Existing approaches with key points descriptors, pixel comparison, or using one before the last output layer of a convolutional neural network do not give such a high-quality result, even with a linear combination of the ones mentioned above. See the embedding builder neural network architecture in Figure 1.
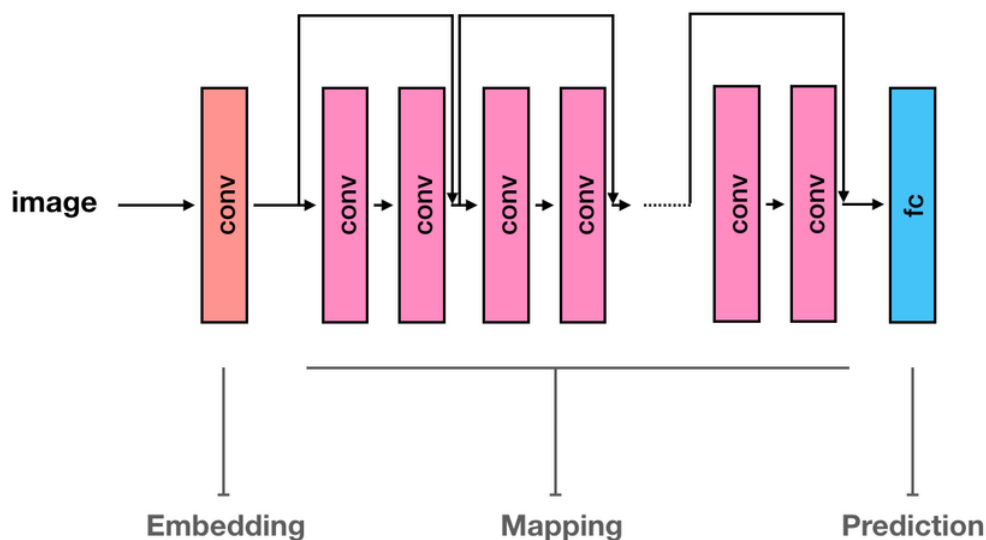

Figure 1: Embedding builder neural network architecture

2.2. Distributed embeddings storage

For some tasks, like multi-million or near-realtime image comparison, storing the image embeddings in distributed storage is essential. Because the generation of the image embedding takes approximately 1-2 seconds on Nvidia 1080Ti GPU, with the next usage in the decision block, there will be no need to re-process the image by the feature extractor. It is proposed to use a document-oriented database since no relation between the compared objects is expected, and storing data in JSON documents is an advantage. Thus, any key-value storage that involves storing massive things as values is suitable for storing feature vectors. We used MongoDB in all our experiments.

2.3. Decision-making unit

The decision-making unit entirely determines the application area of the image embedding. Let's consider two practical applications: finding near duplicate images and clustering rooms based on photos from different shooting angles. This model was applied primarily to the internal private datasets of

images by LUN.UA, which is a country-leading portal in real estate. To solve the proposed problem, collecting a dataset of pairs of sample images is necessary. In the first case, we needed to collect pairs that are considered to be near-duplicates and which are not considered as such. In the second, there are pairs where the photo of the same room is taken from different angles, and various rooms are taken from random view angles.Then we run the embeddings builder and form a vector for each image in the sample dataset. Each pair from the sample forms its new vector, which is obtained by combining the Euclidean metrics, L1 and cosine distance, etc. (*It's allowed to use the proposed metrics, taking into account the equivalence of measures in finite-dimensional spaces. However, there are non-equivalent norms for infinite-dimensional spaces, and using a combination of metrics in the resulting vector can significantly improve the quality of the comparison*.) Image feature vectors often turn out to be sparse, so the additional use of the cosine distance is highly influential.

The new image pair vector formed in this way will be used to train a new fully-connected N-layer neural network with one output neuron. As a result of training, the decision block is trained to compare pairs of images. Depending on the training sample, the block can solve a particular problem.

3. Results, estimations, and the discussion

Building the image embedding in a proposed way and applying it to the task of image near-duplicate detection showed incredible results on the private datasets (LUN.UA real estate images) – 8-10x fewer mistakes in duplicate determination compared to SIFT, SURF, ORB keypoints algorithms. We have proceeded with experiments on the two mentioned sub-types of the image comparison tasks, namely:

- near duplicate images of various graphical contexts (dataset size: 80 000 image pairs, see examples in Figure 2),
- multi-angle photos of the (same) rooms (dataset size: 12 500 image pairs, see examples in Figure 3).

These datasets are private now, but we are working on making these data publicly available. The comparison was done for our solution and 3 alternative techniques:

- image embedding formed by taking previous before the last layer of pre-trained ResNet50 – the image feature vector,
- SIFT / SURF / ORB descriptors,
- perceptual DCT hash,
- image embedding formed by the combination of intermediate layers of ResNet50 (the proposed method, see the scheme in Figure 4).

The Precision, Recall, and F_1 measure values are presented in Table 1 and Table 2 for these two tasks.

Table 1
Near duplicate images of various graphical contexts task & dataset (80 000 image pairs)
The best results are highlighted **in bold**

Method	Precision	Recall	F_1
ResNet50 feature vector	0.82	0.87	0.84
SIFT/SURF/ORB descriptors	0.81	0.84	0.82
Perceptual DCT hash	0.71	0.82	0.76
The proposed method	**0.94**	**0.96**	**0.94**

Table 2
Multi-angle photos of the rooms task & dataset (12 500 image pairs)
The best results are highlighted **in bold**

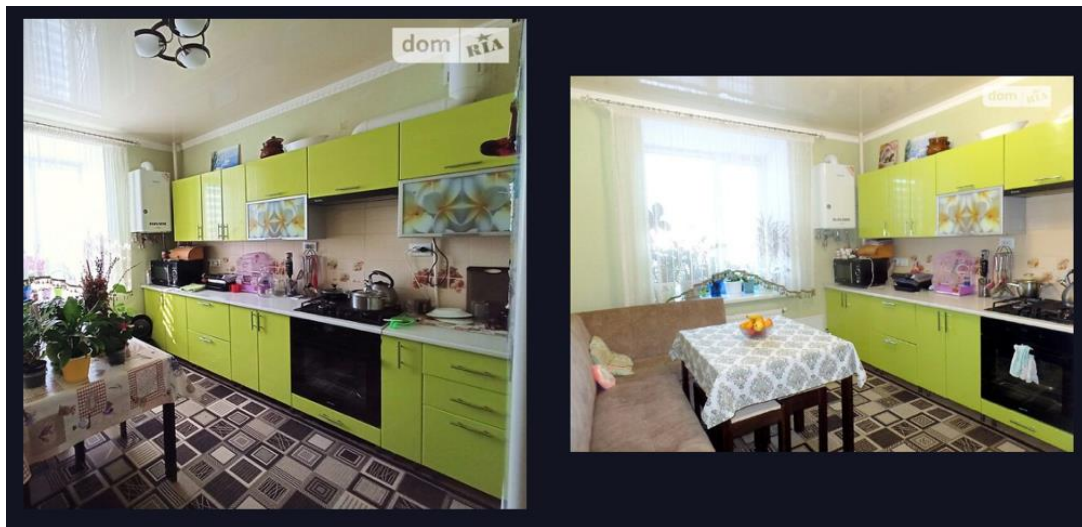| Method	Precision	Recall	F_1
ResNet50 feature vector	0.81	0.79	0.79
SIFT/SURF/ORB descriptors	0.81	0.75	0.77
Perceptual DCT hash	0.62	0.72	0.66
The proposed method	**0.84**	**0.91**	**0.87**

Figure 2: Near duplicate images in the LUN's dataset: the same room on similar photos.


(a)

(b)

Figure 3: Multi-angle photos of the (same) rooms (a), and (b), from the LUN's dataset.

Thus, we can conclude, that the proposed model is precise enough for the task stated, and also the method works quite effectively on the proper hardware (up to 2 sec for images up to 12 megapixels on the NVIDIA 1080Ti GPU chip). Table 1 and 2 shows that the proposed model outperforms other known techniques and shows the best benchmarks in the tests conducted. Also, the model should be tested over other available datasets to ensure generality. Authors are going to do this in future work.
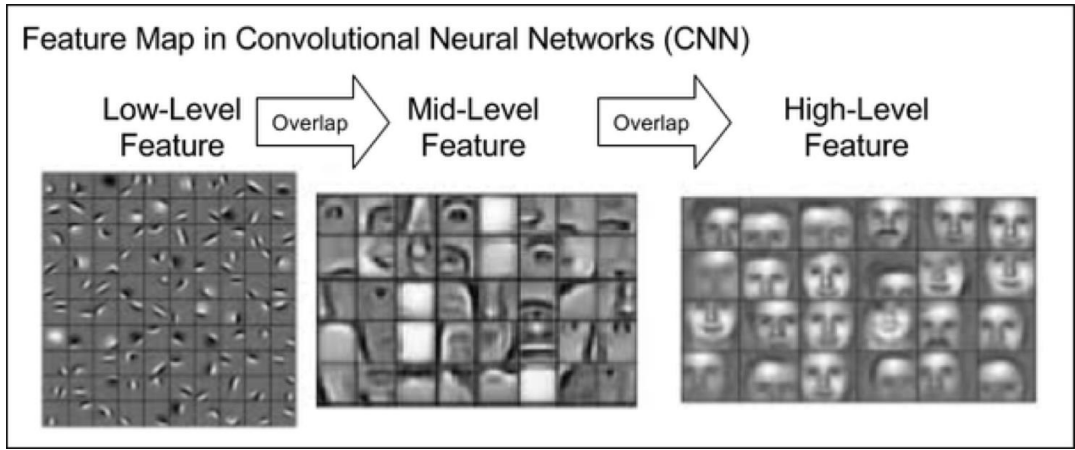
We suggest the proposed model also should be effective for similar tasks mentioned in the Introduction section. This should be checked in the next research.
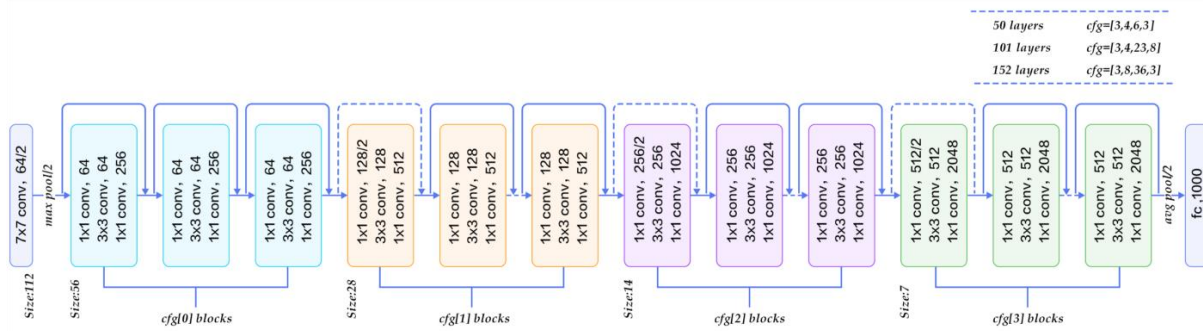
(a)
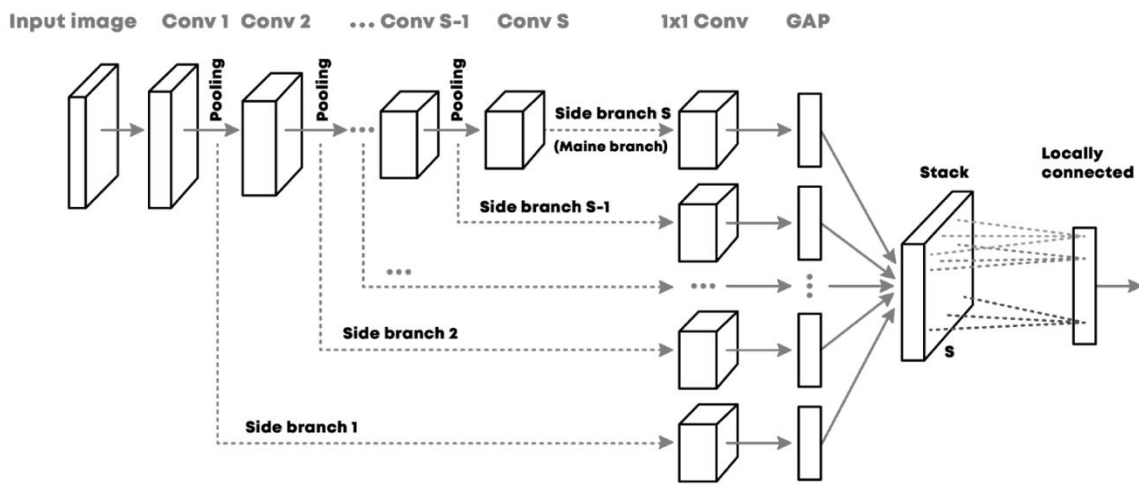
(b)

(c)

Figure 4: The scheme of image embedding construction from the combination of intermediate layers of ResNet50 architecture (a) in the proposed method (b), and the low-, mid-, and high-level features in CNN (c).

4. Further research

We suppose the much broader applications of the model proposed to other tasks and more applications in other spheres for similar tasks. So, further fundamental research is needed on the topic:

- to investigate the influence of the initial architecture of a convolutional neural network, with the layers of which we make a combination to build an image embedding;

- to investigate the influence of pre-training of the selected architecture on image classification tasks per 1000 category because the previous training let us consider the convolutional neural network as a feature extractor;
- to investigate possible options for choosing layers, their number, and the method of combinations (concatenation, averaging, difference) – often called the meta-parameters tuning;
- to investigate the applicability of vector representation for image "compression" (packing into a vector – then transfer – and then unpacking) – to enhance the application possibilities of the model proposed, a kind of transfer learning technique;
- to analyze the effectiveness of application on such classes of tasks as recognition of "stylistics" of images, classification of forged scanned copies of documents, visual navigation, and recognition of diseases by X-ray or MRI images, which would extend the model applicability dramatically.

5. Conclusions

In this work, we proposed a new model for image description – the image embedding vector construction and demonstrated its applications.

The task and its applications were overviewed. The known methods were outlined, namely, the DCT hash approach, SIFT, SURF, ORB methods, key points, CNN, and ResNet-50 as the most promising among them. The model, its inner structure, and the motivation for it were presented here. The main idea of the model proposed is to combine the selected low-level, mid-level and high-level features from the CNN constructed to achieve better precision and F_1 score over the LUN's dataset.

Then this model was tested in an e-commerce task [49-53], and applied to the real-world dataset of LUN.UA, namely, the private set of real estate images, and obtained an excellent result, which exceeds expectations and appeared to be much better than competitors – previously known models and approaches, being estimated by F_1 measure. The benchmarks and calculations supported this conclusion. The promising experimental results demonstrate the validity and effectiveness of the proposed model. Now, this model is in production use in LUN.UA. Its research and development continue. Also, the next questions for further research were highlighted here.

6. References

[1] K. K. Thyagharajan, G. Kalaiarasi, A Review on Near-Duplicate Detection of Images using Computer Vision Techniques, Archives of Computational Methods in Engineering 28.3 (2021): 897–916.
[2] J.J. Foo, R. Sinha, J. Zobel, SICO: a system for detection of near-duplicate images during search, in: 2007 IEEE International Conference on Multimedia and Expo, IEEE (2007): 595–598.
[3] J.J. Foo, J. Zobel, R. Sinha, S.M.M. Tahaghoghi, Detection of near-duplicate images for web search, in: Proceedings of the 6th ACM International Conference on Image and Video Retrieval (2007): 557–564.
[4] X. Wu, A.G. Hauptmann, C.W. Ngo, Practical elimination of near-duplicates from web video search, in: Proceedings of the 15th ACM international conference on Multimedia (2007): 218–227.
[5] W.L. Zhao, X. Wu, C.W. Ngo, On the annotation of web videos by efficient near-duplicate search, IEEE Transactions on Multimedia 12.5 (2010), 12.5: 448–461.
[6] X. Wu, C.W. Ngo, A.G. Hauptmann, H.K. Tan, Real-time near-duplicate elimination for web video search with content and context, IEEE Transactions on Multimedia 11.2 (2009): 196–207.
[7] M. Bhavani, V.A. Narayana, G. Sreevani, A novel approach for detecting near-duplicate web documents by considering images, text, size of the document and domain, in: ICCCE 2020, Springer, Singapore (2021): 1355–1366.
[8] W.T. Chu, C.H. Lin, Consumer photo management and browsing facilitated by near-duplicate detection with feature filtering, Journal of Visual Communication and Image Representation 21.3 (2010): 256–268.

[9] A. Jinda-Apiraksa, V. Vonikakis, S. Winkler, California-ND: An annotated dataset for near-duplicate detection in personal photo collections, in: 2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX), IEEE (2013): 142–147.

[10] J.J. Foo, J. Zobel, R. Sinha, Clustering near-duplicate images in large collections, in: Proceedings of the International Workshop on Multimedia Information Retrieval (MIR'07), Association for Computing Machinery, New York, NY, USA (2007): 21–30.

[11] G. Kalaiarasi, K.K. Thyagharajan, Clustering of near duplicate images using bundled features, Cluster Computing 22.5 (2019): 11997–12007.

[12] Y.G. Jiang, C.W. Ngo, Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval, Computer Vision and Image Understanding 113.3 (2009): 405–414.

[13] S. Liang, P. Wang. An efficient hierarchical near-duplicate video detection algorithm based on deep semantic features, in: International Conference on Multimedia Modeling, Springer, Cham (2020): 752–763.

[14] D. Liu, J. Shen, A. Wang, C. Wang, Secure real-time image protection scheme with near-duplicate detection in cloud computing, Journal of Real-Time Image Processing 17.1 (2020): 175–184.

[15] G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, I. Kompatsiaris, Finding near-duplicate videos in large-scale collections, in: Video Verification in the Fake News Era, Springer, Cham (2019): 91–126.

[16] W. Dong, Z. Wang, M. Charikar, K. Li, High-confidence near-duplicate image detection, in: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval (2012): 1–8.

[17] Z. Zhou, Q.J. Wu, F. Huang, X. Sun, Fast and accurate near-duplicate image elimination for visual sensor networks, International Journal of Distributed Sensor Networks 13.2 (2017): 12p.

[18] T.E. Koker, S.S. Chintapalli, S. Wang, B.A. Talbot, D. Wainstock, M. Cicconet, M.C. Walsh, On Identification and Retrieval of Near-Duplicate Biological Images: a New Dataset and Protocol, in: 2020 the 25th International Conference on Pattern Recognition, ICPR, IEEE (2021): 3114–3121.

[19] S. Hadipour, S. Aram, R. Sadeghian, Similar multi-modal image detection in multi-source dermatoscopic images of cancerous pigmented skin lesions, in: Advances in Computer Vision and Computational Biology, Springer, Cham (2021): 109–119.

[20] S. Srivastava, P. Mukherjee, B. Lall, imPlag: Detecting image plagiarism using hierarchical near duplicate retrieval, in: 2015 Annual IEEE India Conference (INDICON), IEEE (2015): 1–6.

[21] B. Mehta, S. Nangia, M. Gupta, W. Nejdl, Detecting image spam using visual features and near duplicate detection, in: Proceedings of the 17th international conference on World Wide Web (2008): 497–506.

[22] Z. Wang, W.K. Josephson, Q. Lv, M. Charikar, K. Li, Filtering image spam with near-duplicate detection, in: CEAS (2007).

[23] Y. Ke, R. Sukthankar, L. Huston, Efficient near-duplicate detection and sub-image retrieval, ACM Multimedia 4.1 (2004): 5p.

[24] Y. Ke, R. Sukthankar, L. Huston, An efficient parts-based near-duplicate and sub-image retrieval system, in: Proceedings of the 12th annual ACM International Conference on Multimedia (2004): 869–876.

[25] F. Nian, T. Li, X. Wu, Q. Gao, F. Li, Efficient near-duplicate image detection with a local-based binary representation, Multimedia Tools and Applications 75.5 (2016): 2435-2452.

[26] W.L. Zhao, C.W. Ngo, H.K. Tan, X. Wu, Near-duplicate keyframe identification with interest point matching and pattern learning, IEEE Transactions on Multimedia 9.5 (2007): 1037–1048.

[27] W.L. Zhao, C.W. Ngo, Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection, IEEE Transactions on Image Processing 18.2 (2009): 412–423.

[28] Y. Lei, L. Zheng, J. Huang, Geometric invariant features in the Radon transform domain for near-duplicate image detection, Pattern Recognition 47.11 (2014): 3630–3640.

[29] Y. Li, A Fast Algorithm for Near-Duplicate Image Detection, in: 2021 IEEE International Conference on Artificial Intelligence and Industrial Design, AIID, IEEE (2021): 360-363.

[30] Q. Chum, J. Philbin, A. Zisserman, Near duplicate image detection: Min-hash and TF-IDF weighting, in: BMVC 810 (2008): 812–815.

[31] L. Liu, Y. Lu, C.Y. Suen, Variable-length signature for near-duplicate image matching, IEEE Transactions on Image Processing 24.4 (2015): 1282–1296.

[32] L. Xie, Q. Tian, W. Zhou, B. Zhang, Fast and accurate near-duplicate image search with affinity propagation on the ImageWeb, Computer Vision and Image Understanding 124 (2014): 31–41.

[33] Z. Zhou, K. Lin, Y. Cao, C.N. Yang, Y. Liu, Near-duplicate image detection system using coarse-to-fine matching scheme based on global and local CNN features, Mathematics, 8.4 (2020): 644.

[34] G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, Y. Kompatsiaris, Near-duplicate video retrieval by aggregating intermediate CNN layers, in: International Conference on Multimedia Modeling, Springer, Cham (2017): 251–263.

[35] Y. Zhang, S. Zhang, Y. Li, Y. Zhang, Single- and cross-modality near duplicate image pairs detection via spatial transformer comparing CNN, Sensors, 21.1 (2021): 255.

[36] O. Chum, J. Philbin, M. Isard, A. Zisserman, Scalable near identical image and shot detection, in: Proceedings of the 6th ACM International Conference on Image and Video Retrieval (2007): 549–556.

[37] B. Barz, J. Denzler, Do we train on test data? Purging cifar of near-duplicates, Journal of Imaging, 6.6 (2020): 41.

[38] H. Matatov, M. Naaman, O. Amir, Dataset and case studies for visual near-duplicates detection in the context of social media, arXiv preprint (2022) arXiv:2203.07167.

[39] Dijana Tralic, Iran Zupancic, Sonja Grgic, Mislar Grgic, CoMoFoD: New database for copy-move forgery detection, in: Proceedings of the 55th International Symposium ELMAR (2013): 49–54.

[40] L. Morra, F. Lamberti, Benchmarking unsupervised near-duplicate image detection, Expert Systems with Applications 135 (2019): 313–326.

[41] Bjorn Barz and Joachim Denzler, Hierarchy-based Image Embeddings for Semantic Image Retrieval (2019) URL: https://arxiv.org/pdf/1809.09924v4.pdf

[42] Maxim Berman, Herve J´egou, Andrea Vedaldi, Iasonas Kokkinos and Matthijs Douze, MultiGrain: a unified image embedding for classes and instances (2019) URL: https://arxiv.org/pdf/1902.05509v2.pdf

[43] Zehao Yu, Jia Zheng, Dongze Lian, Zihan Zhou and Shenghua Gao, Single-Image Piece-wise Planar 3D Reconstruction via Associative Embedding (2019) URL: https://arxiv.org/pdf/1902.09777v3.pdf

[44] Anita Rau, Guillermo Garcia-Hernando, Danail Stoyanov, Gabriel J. Brostow and Daniyar Turmukhambetov, Predicting Visual Overlap of Images Through Interpretable Non-Metric Box Embeddings (2020) URL: https://arxiv.org/pdf/2008.05785v1.pdf

[45] Guang Feng, Zhiwei Hu, Lihe Zhang and Huchuan Lu, Encoder Fusion Network with Co-Attention Embedding for Referring Image Segmentation (2021) URL: https://arxiv.org/pdf/2105.01839v1.pdf

[46] Maryam Asadi-Aghbolaghi, Reza Azad, Mahmood Fathy and Sergio Escalera, Multi-level Context Gating of Embedded Collective Knowledge for Medical Image Segmentation (2020) URL: https://arxiv.org/pdf/2003.05056v1.pdf

[47] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, Deep Residual Learning for Image Recognition (2015) URL: https://doi.org/10.48550/arXiv.1512.03385

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, IEEE Computer Vision and Pattern Recognition, CVPR (2009).

[49] V.L. Pleskach, E-commerce technologies, Kyiv, KNTEU (2004): 226p.

[50] T.I. Lytvynenko, T.V. Panchenko and V.D. Redko, Sales Forecasting using Data Mining Methods, Bulletin of Taras Shevchenko National University of Kyiv, Series: physical-mathematical sciences 4 (2015): 148–155.

[51] I. Bieda, T. Panchenko, A Systematic Mapping Study on Artificial Intelligence Tools Used in Video Editing, International Journal of Computer Science and Network Security 22.3 (2022): 312–318.

[52] I. Bieda, A. Kisil, T. Panchenko, An Approach to Scene Change Detection, in: Proceedings of the 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2021) volume 1 (2021): 489–493.

[53] I. Bieda, T. Panchenko, A Comparison of Scene Change Localization Methods over the Open Video Scene Detection Dataset, International Journal of Computer Science and Network Security 22.6 (2022): 1–6.