This is a post-peer-review, pre-copyedit version of a paper published in Mansouri A, El Moataz A, Nouboud F & Mammass D (eds.) *Image and Signal Processing*. Lecture Notes in Computer Science, 10884. ICISP 2018: International Conference on Image and Signal Processing, Cherbourg, France, 02.07.2018-04.07.2018. Cham, Switzerland: Springer International Publishing, pp. 284-291. The final authenticated version is available online at: https://doi.org/10.1007/978-3-319-94211-7_31

Editorial Image Retrieval using Handcrafted and CNN Features

Claudia Companioni-Brito[1], Mohamed Elawady[2], Sule Yildirim[1], and Jon Yngve Hardeberg[1]

[1] NTNU - Norwegian University of Science and Technology, Gjøvik, Norway
[2] Université de Lyon, UJM-Saint-Etienne, CNRS, IOGS, Laboratoire Hubert Curien UMR5516, F-42023, Saint-Etienne, France
sule.yildirim@ntnu.no

Abstract. Textual keywords have been used in the early stages for image retrieval systems. Due to the huge increase of image content, an image is efficiently used instead according to the time computation. Deciding powerful feature representations are the important factors for the retrieval performance of a content-based image retrieval (CBIR) system. In this work, we present a combined feature representation based on handcrafted and deep approaches, to categorize editorial images into six classes (athletics, football, indoor, outdoor, portrait, ski). The experimental results show the superior performance of the combined features among different editorial classes.

Keywords: Image features, Similarity, CBIR, CNN, LBP, BoVW

1 Introduction

Content-based image retrieval (CBIR) is one of the important research challenges employed in computer vision field during the last decades [1]. The main goal of CBIR is searching the database images by investigating the correct representation of their visual contents. Different feature descriptors have been proposed for image representation [2], ranging from global features based on color and textural information, then local feature representations (i.e bag-of-words models using local descriptors), into recent deep CNN features presenting a generic semantic presentation to understand the image scene. Based on editorial images, the idea to use the query image as an input for the proposed CBIR framework, as shown in Figure 1, to retrieve similar images. This work makes the following contributions: (1) constructing an editorial database for image retrieval purpose, (2) introducing a combined feature-based framework for CBIR by conducting reliable feature representations of images.

The rest of this paper is organized as follows: Section 2 presents the state-of-the-art for CBIR systems using hand-features and deep features. Section 3 introduces the CBIR framework based on the combined feature representation. Section 4 presents the details of editorial dataset and the experimental information. In addition, the qualitative and quantitative results are discussed against

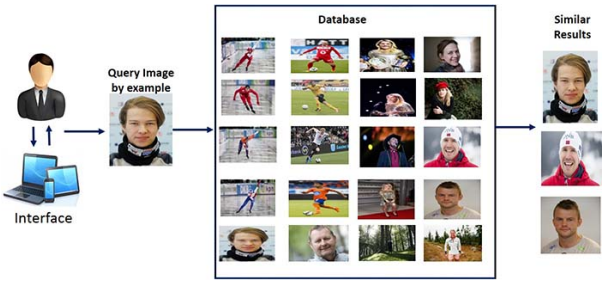


Fig. 1: An illustration of the query scheme within the corresponding retrieval results.

the proposed feature-based approaches among different editorial queries. Finally, Section 5 contains the conclusion and future work.

2 Related work

Content-based image retrieval (CBIR) is one of the main research challenges in the field of computer vision. Traditionally CBIR uses low-level feature descriptors over the recent years, provided that these descriptors have been developed for describing the images at a global scale, such as color features [3,4], texture features [3] and shape features [4]. Bag of Words (BoW) [5] model using local level feature descriptors (i.e. SIFT [6] and LBP [7]) has also been extensively explored as a routine image representation on CBIR [8]. The descriptors describe the image patches of an object, while BoW outputs global features up on the generation of a visual-word vocabulary based on these local descriptors. These features are still not solve the problem of the semantic gap between low level features and the visual features of human perception. This is essentially due to the fact that current image representation schemes are hand-crafted and insufficient to capture the semantics [8]. The advances of machine learning algorithms have provided a new way of reducing the semantic gap.

Since deep learning has been developed, is one of the hot topics in computer vision nowadays, and has been successfully applied to many digital applications e.g. image classification [9], visual tracking [10], and CBIR [2,11,12]. Deep Convolutional Neural Network (CNN) has been leading position on feature extraction and representation with the use of deep features for image retrieval. Based on CNN architecture, several global descriptors have been proposed to use pre-trained or learned networks. Babenko et al. [13] investigated the impact of the retrieval performance of the corresponding neural codes, retraining CNN on different datasets. The advantages of transfer learning of generic CNN features, trained on very large classification image datasets, to be used for image retrieval has shown a noticeable performance over training datasets from scratch requiring manual effort. Wan et al. [2] applied many existing deep learning methods for learning feature representation from images with application to CBIR tasks.

Also [13] and then confirmed by [11], showed that pre-trained models on ImageNet for object classification could be improved by fine-tuning on an external set of Landmarks images. [11] used a similar CNN architecture proposed for [9], first training a deep learning model from a large collection of training data; then applying the trained deep model for learning feature representations in a new domain.

Recent works [12] introduced a new deep bilinear CNN architecture in the context of visual CBIR, which reduced the dimension of the extracted features into a low-dimensional image representation using a modified pooling scheme of the compact bilinear pooling. [14] carried out a comparison of deep convolutional features and SIFT. It investigated some possible ways to aggregate local deep features to produce compact global descriptors, and suggesting a new descriptor based on a simple sum-pooling aggregation with the best results for image retrieval. Addressing the semantic gap between low-level features and high-level semantic features, [15] proposed image retrieval using fused deep features from two different deep convolutional networks: LeNet-5 and AlexNet.

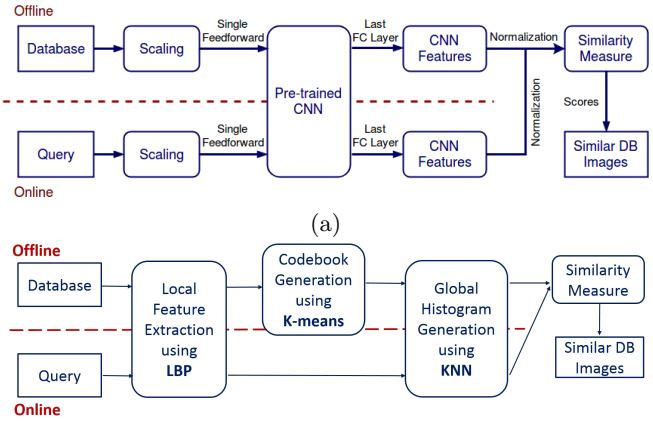


Fig. 2: The general framework of deep (a) and hand-crafted (b) features extraction. The modules above and below the red dashed line are in the off-line stage and on-line stage, respectively.

This work is based on the study of the convolutional neural network for editorial image retrieval. Deep features are extract from three existing CNNs, and fusing to get the image feature information. Hand-crafted features are also fuse to improve the retrieval results.

3 Methodology

Based on the idea of deep transfer learning [16], pre-trained CNN models on the image classification task of ImageNet ILSVRC 2013 (over 1.2 million images,

1,000 classes) used the vector representation of some fully-connected layers as powerful features, producing superior results compared to the state-of-the-art methods on different applications of visual recognition tasks (i.e. image classification and retrieval [13]). As shown in Figure 2a, the collection of images in the database are scaled into the required size of network constraints, and then passed in a single scan though pre-trained CNN models. Afterwards, the deep features are extracted from the last fully connected layer, and then normalized in the advantage of domain independence cause [16]. These features are linked with each corresponding database image in the offline mode, for saving the computation time in CBIR systems. In the online mode, the query image is selected to extract and normalize the corresponding feature in the same CNN network. The similarity process is executed between the features of query and database to find the best similarity images, respect to the query image. The deep features are constructed through the concatenation of the fully-connected layers from a set of different CNN models (VGG-VD [17], GoogLeNet [18], ResNet [19]). Such models are pre-trained on ImageNet and can be used as generic feature extractors for other tasks [16,20].

Bag-of-Words [5] along side with local feature extraction methods (i.e. SIFT, LBP, HOG) have been introduced in image retrieval systems [1] producing reliable accuracy results. As shown in Figure 2b, local features are extracted from the database images in offline mode using LBP [7]. Histogram-based codebook are generated using K-means clustering algorithm respect to the extracted features, followed by generation of the global histogram-based representations using KNN search method. These histograms are attached with corresponding database images for faster computation in the online mode. Given a query image, its global representation is computed using the offline codebook, and compared with the database features for top similarity retrieval results.

4 Results and Discussion

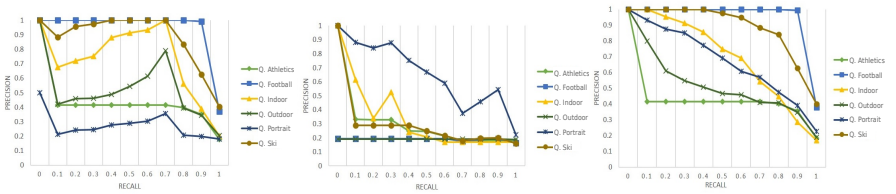
This work has an industrial collaboration with NTB company for providing an editorial database: 17,000 medium-resolution images. Around 1,000 images have been randomly selected from the database, and divided into 6 categories ($\sim$ 160 for each category): Athletics, Football, Indoor-Night, Outdoor-Daytime, Portrait, Ski. Figure 3 shows some sample images of selected categories from NTB database.

Table 1: Details of deep and handcrafted features used in the proposed work

Method Name (Year)	Feature Size
LBP (1996) [7]	500 (BoW [5])
VGG-VD (2014) [17]	4096 (Last FC)
GoogLeNet (2015) [18]	1024 (Last FC)
ResNet (2016) [19]	2048 (Last FC)

Fig. 3: Samples images of the NTB database. From column one to six: Athletics, Football, Outdoor-Daytime, Indoor-Night, Portrait, Ski images respectively.

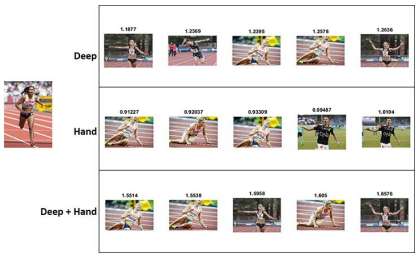
The proposed work is implemented using MATLAB (R2016b) on windows-based PC platform along side with research-based libraries: *MatConvNet* for deep learning models, and *VLFeat* for image processing functions. the used cell size for LBP feature extraction set to 8×8. Table 1 presents the dimensional information of the deep and handcrafted features used in this work. Afterwards, these features are normalized using $l2$ norm. The combined feature can be constructed by concatenating the proposed feature vectors. Euclidean distance is selected as a similarity measure between database and query features.



(a) Deep (b) Hand (c) Deep+Hand

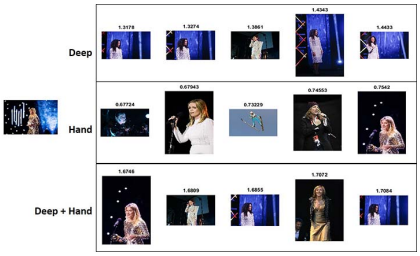
Fig. 4: Precision and recall curves of deep, handcrafted and combined features extracted for six categories. Combining deep and handcrafted features include more comprehensive and effective image information. Results of Precision-Recall curves are improved in some cases (e.g. Q. Portrait).

The performance of our retrieved system is measured using precision-recall curves showed in Figure 4. In comparison between the deep features and hand-



(a) QueryA: Athletics

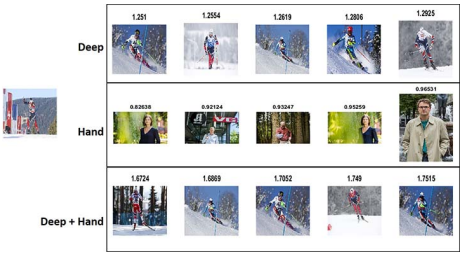
(b) QueryF: Football



(c) QueryI: Indoor-Night

(d) QueryO: Outdoor-Daytime

(e) QueryP: Portrait



(f) QueryS: Ski

Fig. 5: Retrieval results of top 5 similar images for the following categories: Athletics, Football, Indoor-Night, Outdoor, Portrait, Ski. For each category, the query image appears in the left, while the retrieval results display in 3 rows representing different approaches (deep features using pre-trained CNN models 'Deep', handcrafted features using LBP with BoVW 'Hand', fused features of deep and handcrafted features 'Deep+Hand'). Similarity scores state above the retrieved images.

crafted features, the best precision among the most recall values is located in the deep curve using the football query and (portrait, ski, indoor-night) queries in the handcrafted curve. The worst precision rate is portrait query in the deep curve, while all the queries except portrait have bad precision results in the handcrafted curve. The combined features get the advantage of having the best precision values for (football, ski) queries from the deep curves and portrait query from the handcrafted curve, while the other queries (indoor-night, outdoor-daytime, athletics) get intermediate precision results.

Some examples of the top 5 similar images from the database are retrieved in Figure 5, respect to each query image from six categories (Athletics, Football, Indoor-Night, Outdoor-Daytime, Portrait, Ski). In athletics query, deep and combined features get all top 5 retrieved results as correctly matched within the query, while the handcrafted features got two incorrect (football) results in 4th and 5th retrieved images. The combined features have advantages in first two retrieved images by showing the player with the same uniform colors plus the yellow object in the background. In football query, all different types of the proposed features get all correct results among 5 retrieval images. In indoor-night query, the deep features have similar results among top 5 images, while the handcrafted features ranked top 3 image from a different category (ski). The combined features have the same girl within the same event in top 1 retrieval image. In outdoor-daytime query, the deep features got incorrect category (portrait) in top 4 results, while the handcrafted and combined features have all correct results. It summarizes that deep features describe well the global context of the scene, while the handcrafted features have better use representing the object details inside an image. In portrait query, all different types of the proposed features matched the results right within the same category except in top 2 retrieval image (indoor-night) of handcrafted features. In last query (Ski), the deep and combined features have correct category among top 5 images, while the handcrafted features got all retrieval image from different category (outdoor-daytime).

5 Conclusion

This work showed an experimental study for different feature extraction approaches for editorial image retrieval systems. The results show that the deep features describe well the global context of image scene, while the handcrafted features are better in representing the intra-object details inside an image. The fused features of both deep and handcrafted approaches have the best retrieval results. In future work, efficient dimensional reduction will be applied to the deep features without losing-out similar accuracy results. Plus, a reliable multiple dictionary method for bag-of-words can be introduced to fuse different hand-crafted features (i.e. SIFT, HOG, LBP).

Acknowledgments

I would like to thank NTB, the Norwegian news agency, for providing the dataset for research use, from the Scanpix database.

References

1. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **2**(1) (2006) 1–19

2. Wan, J., Wang, D., Hoi, S.C.H., Wu, P., Zhu, J., Zhang, Y., Li, J.: Deep learning for content-based image retrieval: A comprehensive study. In: Proceedings of the 22nd ACM international conference on Multimedia, ACM (2014) 157–166
3. Yue, J., Li, Z., Liu, L., Fu, Z.: Content-based image retrieval using color and texture fused features. Mathematical and Computer Modelling **54**(3-4) (2011) 1121–1127
4. Jain, A.K., Vailaya, A.: Image retrieval using color and shape. Pattern recognition **29**(8) (1996) 1233–1244
5. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 2., IEEE (2005) 524–531
6. Lowe, D.G.: Object recognition from local scale-invariant features. In: Computer vision, 1999. The proceedings of the seventh IEEE international conference on. Volume 2., Ieee (1999) 1150–1157
7. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern recognition **29**(1) (1996) 51–59
8. Zhou, W., Li, H., Tian, Q.: Recent advance in content-based image retrieval: A literature survey. arXiv preprint arXiv:1706.06064 (2017)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
10. Wu, G., Lu, W., Gao, G., Zhao, C., Liu, J.: Regional deep learning model for visual tracking. Neurocomputing **175** (2016) 310–323
11. Gordo, A., Almazán, J., Revaud, J., Larlus, D.: Deep image retrieval: Learning global representations for image search. In: European Conference on Computer Vision, Springer (2016) 241–257
12. Alzu'bi, A., Amira, A., Ramzan, N.: Content-based image retrieval with compact deep convolutional features. Neurocomputing **249** (2017) 95–105
13. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. In: European conference on computer vision, Springer (2014) 584–599
14. Babenko, A., Lempitsky, V.: Aggregating local deep features for image retrieval. In: Proceedings of the IEEE international conference on computer vision. (2015) 1269–1277
15. Liu, H., Li, B., Lv, X., Huang, Y.: Image retrieval using fused deep convolutional features. Procedia Computer Science **107** (2017) 749–754
16. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. (2014) 806–813
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
18. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 1–9
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
20. Athiwaratkun, B., Kang, K.: Feature representation in convolutional neural networks. arXiv preprint arXiv:1507.02313 (2015)