International Journal of Multimedia Information Retrieval (2023) 12:36
https://doi.org/10.1007/s13735-023-00304-6

EDITORIAL

Special Issue on Open-Domain Image Retrieval in the Wild

Yu Liu[1] · Yanming Guo[2] · Yusuke Matsui[3]

Published online: 8 November 2023

We are currently in an era of abundant image data. With a surge in real-world image samples, efficient image retrieval methods are urgently needed. However, existing content-based image retrieval largely focuses on predefined classes and single-task training in closed-domain scenarios. Yet, in open-domain applications like online shopping recommendations and security-related person re-identification, data distributions are complex and unpredictable. This presents a major challenge and new research avenues for open-domain image retrieval. While numerous image retrieval works have been presented in top conferences and journals, a dedicated special issue consolidating mature research in open-domain image retrieval is yet to be established. This issue serves as a platform for researchers and practitioners, especially in multimedia and computer vision, to showcase high-quality research and encourage discussions on future directions in this critical research area.

This special issue aims to collate state-of-the-art research focusing on open-domain image retrieval. The objectives include models capable of: (1) retrieving novel and unseen classes with zero-shot transfer; (2) continual learning of a sequence of new tasks; (3) performing multimodal retrieval across images and other modalities (e.g., sketch, text, audio) for applications like fashion recommendation, biometrics, and social media. We received 22 submissions, of which 12 papers were selected for publication after at least a double peer-review process. We are pleased to present them in the following.

For the task of few-shot image classification on the vision-language model, the paper by Jie Yan, Yuxiang Xie, Yanming Guo, Yingmei Wei, Xiaoping Zhang, and Xidao Luan, presents an input-specific neural network to alleviate the overfitting issue. Besides, this paper introduces learned visual features into the pre-trained CLIP model as prompts to perform automatic prompt tuning and realize the learning of the input of the specified cue vector and the establishment of connections between vision and text. Two learnable lightweight neural networks are added at the end of CLIP to guide the transmission of information between different classes through fine-tuning visual and textual features.

Toward the recognition and detection of multi-oriented text from textual natural scene images, in the paper by Shilpa Mahajan, Rajneesh Rani, Karan Trehan, the authors produce a largest scene text dataset NITJ-WS and introduce a new deep and lightweight architecture to segment the text block image at the word level, which confirms their effectiveness contrast with UNet and ResUNet.

Image retrieval and recommendation can be applied in multiple downstream tasks, e.g., online shopping. For instance, the paper by Ahmad Alzu'bi, Lojin Bani Younis, and Alia Madain proposes a novel approach that enables consumers to choose the right outfit and virtually try it on by uploading a frontal image of their entire body and generating a 3D fitting image. The approach adopts OpenPose for estimating the user's pose, M3D-VTON for matching the selected clothing image to the user' s body, and ADDE for fashion retrieval and recommendation.

For the task of facial expression recognition, the paper by Faten Khemakhem and Hela Ltif introduces NST-GAN to remove the influence of identity-related features on face expression recognition and detect expression information by synthesizing identity. The excellent performance has been determined with three publicly available datasets.

Modeling human behavior and activity patterns to recognize particular events has aroused wide interest in recent years. Specifically, the paper by Mohd. Aquib Ansari, Dushyant Kumar Singh, and Vibhav Prakash Singh presents an advanced three-dimensional convolutional network, which contains a fifteen layers deep architecture, to determine the abnormal human acts in megastores. In addition, the authors create a novel dataset consisting of numerous video clips to represent abnormal behaviors of humans in megastores or shops.

✉ Yu Liu
liuyu8824@gmail.com

[1] Dalian University of Technology, Dalian, China

[2] Hunan Institute of Advanced Technology, Changsha, China

[3] The University of Tokyo, Tokyo, Japan

Datasets have a crucial impact on task research. Therefore, the paper by Sk Maidul Islam and Subhankar Joardar considers the lack of suitable jewelry datasets. Thus, they created a novel ornament dataset OrnamentFIR, which includes over 4.4 K high-quality images of bangles, over 4.8 K high-definition images of necklaces, and more than 2.6 K high-quality images of earrings. A matching network is then employed to extract desired images in the appropriate category from the dataset.

Cross-modal tasks have been developed for decades. In the paper by Mingyue Liu, Honggang Zhao, Longfei Ma, and Mingyong Li, the authors propose a novel decoding context optimization to optimize context prompts for simulating better image–text interaction. Extensive experiments have been performed to verify the effectiveness of this model on eleven image classification datasets.

Toward student engagement, the paper, by Sandeep Mandia, Kuldeep Singh, and Rajendra Mitharwa, presents a student engagement estimation method in the authentic classroom environment, which utilizes a graph convolution network to extract more contributing features. Besides, a learning-centered affective state dataset is curated from existing open source. Multiple ablation studies and experiments have determined the performance outperformed existing state-of-the-art methods.

For the task of cross-modal retrieval, the paper by Xiaohan Yang, Zhen Wang, Wenhao Liu, Xinyi Chang, and Nannan Wu, proposes a deep adversarial multi-label cross-modal hashing algorithm, which employs multi-label and deep feature to establish modal neighbor matrix. Furthermore, the authors design linear classifiers to predict semantic labels for binary features and minimize the hash semantic retention loss to make it have the same semantic information as the sample label. Four loss functions are conducted to guarantee the excellent performance of this task.

The research on medical image watermarking has strong practicability and research value. The paper by Shehu Ayuba and Wan Mohd Nazmee Wan Zainon presents trends in the application of watermarking research in medical images and evaluates the methods recently adopted by researchers. In addition, the survey assesses existing work that meets standard benchmarks in terms of design and performance and discusses other possible research opportunities in the field of medical image watermarking.

Deep neural networks with pyramids have been at the center of attention for image retrieval, due to their remarkable role in extracting the content and semantic features of the image. The paper by Fatemeh Taheri, Kambiz Rahbar and Ziaeddin Beheshtifard extracts manual features including color and texture from the semantic pyramid of a deep neural network. The semantic pyramid is the result of feature mapping fusion of different levels in DNN. In addition, the interpretability of eigenvectors is also considered. Besides, t-SNE technique has been used to explain the discriminability of feature vectors between database classes. Besides, they introduce contour criterion and use feature vectors to study the compatibility of intra-class data sets and the resolvability of inter-class datasets.

Convolutional neural networks provide important assistance to the needs of interdisciplinary biomedicine. In the paper by K. Muthureka, Dr U.Srinivasulu Reddy, and Dr. B. Janet, the authors introduce a customed CNN model for handwritten digit recognition on their proposed CP handwritten digit dataset, which has been data cleaned and created class labels by adopting advantage of disproportionate stratified sampling to generalize dataset. Evaluation shows their superiority over state-of-the-art methods.