 DOI: 10.15203/99106-150-2-34

Enhancing Historical Image Retrieval with Compositional Cues

Tingyu Lin, Robert Sablatnig
Computer Vision Lab, TU Wien
1040 Vienna, Austria
{tylin, sab}@cvl.tuwien.ac.at

Abstract

In analyzing vast amounts of digitally stored historical image data, existing content-based retrieval methods often overlook significant non-semantic information, limiting their effectiveness for flexible exploration across varied themes. To broaden the applicability of image retrieval methods for diverse purposes and uncover more general patterns, we innovatively introduce a crucial factor from computational aesthetics, namely image composition, into this topic. By explicitly integrating composition-related information extracted by CNN into the designed retrieval model, our method considers both the image's composition rules and semantic information. Qualitative and quantitative experiments demonstrate that the image retrieval network guided by composition information outperforms those relying solely on content information, facilitating the identification of images in databases closer to the target image in human perception. Please visit https://github.com/linty5/CCBIR to try our codes.

1 Introduction

With the advancements in digital technology and the increasing emphasis on preserving historical records, a wealth of cultural heritage materials has undergone digitization. Paper-based repositories have undergone expansion and transformation into digital repositories, thus providing foundational resources for humanities scholars and fostering the development of related analytical techniques [4], [24], [15]. Consequently, numerous historical image databases and datasets have come into existence. However, manually searching for similar images in extensive multimedia archives is impractical, making automated image and video retrieval systems crucial for accessing and analyzing historical image archives [21]. Existing popular image retrieval methods, primarily based on semantic or content-based matching, often utilize deep learning techniques to enhance feature extraction, focusing on multimodal retrieval and improvements in retrieval speed and security [7], [29]. Meanwhile, image retrieval regarding photographic settings that determine the quality and style of photographs has received much less attention, resulting in the underutilization of information beyond semantics.

One frequently neglected yet vital aspect is compositional cues. Humans quickly recognize composition rules, which are extensively utilized by photographers [20]. Composition rules aid in conveying structural information, thus enhancing the effectiveness of image retrieval [16]. Within photography, the quality assessment includes physical image parameters, such as size, aspect ratio, color depth, and higher-level perceptual aspects. This latter category, computational aesthetics, deals with the rules of composition, color, clarity, and semantics, offering a more nuanced understanding of image quality [5], [6]. Studies on aesthetic features for similarity detection have focused on color composition [10] and sketch-based composition cues [23]. The grayscale nature of most historical images, lacking color information, emphasizes the importance of compositional cues in historical image retrieval. More research is needed to directly extract composition features from images, particularly historical

The First Austrian Symposium on AI, Robotics, and Vision (AIROV24).

ones, for matching purposes. Combining semantic content and composition for retrieval would better assist historians and photography experts in analyzing historical materials, revealing the quality and potential intentions behind photos and films from various perspectives.

To tackle this challenge, we introduce a novel image retrieval approach that synergizes composition and content features, augmented by a specialized training and evaluation pipeline leveraging historical footage. Our method consists of two primary components: a composition feature extraction network and a content retrieval network. The composition network named Composition Clues Network (CCNet) is inspired by the composition branch from C. Hong et al. [13]'s Composition-Aware Cropping Network (CACNet), which assumes that composition rules can be learned and explicitly modeled within the network to guide image cropping effectively. We optimized this composition branch to serve as our image composition information extractor. Trained on the KU-PCP dataset [16], it extracts Class Activation Maps (CAMs) [31] to encode the Key Composition Map (KCM), which is then passed to the image retrieval network to guide the training and retrieval of content features.

Our proposed Content-Based Image Retrieval Network (CBIRNet) merges composition information with content feature extraction. It was trained and tested using selected images and annotations from the publicly available HISTORIAN dataset [11], a richly annotated historical video collection that offers annotations for over-scanned areas, start and end times of shots in videos, and shot types. Experimental results demonstrate that our CBIRNet, leveraging both composition and content information, can find images that are perceptually closer to the target image across various styles compared to networks relying solely on content-based retrieval.

2 Related Work

This section outlines the developmental trajectory and influential approaches in image composition analysis and content-based image retrieval.

2.1 Image Composition Analysis

In photography, image composition refers to the arrangement of elements within the frame, guiding the viewer's attention to the photographer's intended focus. Following artistic principles, this arrangement signifies the harmony of visual elements and is considered a critical factor in assessing aesthetic quality [16], [19]. Current methods for identifying composition rules involve computational feature design, such as calculating the distance between the subject's center and four centroids to detect the rule of thirds, measuring the dominance of diagonal lines, and assessing visual weight balance by the areas of two regions in the golden ratio, among other complex features [2], [25].

While image composition analysis involves subjective elements, it can still be systematically categorized using established photographic principles. Considerable research, especially using CNNs, has been devoted to composition classification. For instance, T. Lee et al. [16]'s work combines CNN-extracted features with composition rules and includes a sky detector for photographic composition classification and element detection. Similarly, C. Hong et al. [13] introduced a composition branch utilizing CNN features and CAMs to form KCM, providing composition information for downstream tasks. Previous efforts have explored conventional and CNN-based approaches within computational aesthetics, with our work further refining these methods and expanding their application domains.

2.2 Content-Based Image Retrieval

Content-Based Image Retrieval (CBIR) is generally categorized into two main tasks: Category-level Image Retrieval (CIR) and Instance-level Image Retrieval (IIR). CIR aims to retrieve images in the same category as the query image, while IIR targets images with the exact instance depicted in the query [3]. This paper concentrates on retrieving historical images that present objects of the same category with a similar layout. The typical CBIR system process is segmented into online and offline stages: "online" refers to operations on the query target image, while "offline" refers to processing the database for the query. Once a target image is an input, its feature vector is extracted and then searched, scored, and ranked against feature vectors from the database images. The results are finally returned and reordered according to similarity [17]. We primarily focus on image representation within this workflow, namely extracting image features.

In CBIR systems, feature extraction methods mainly include conventional and CNN-based approaches. Conventional strategies are further classified as global and local feature extraction. Global attributes encompass color, shape, texture, and structure, allowing combinations across different features [18], [27], [26]. Local features are identified using renowned algorithms such as SIFT and its variations [32], [30], along with codeword-based methods designed for database search efficiency, including Bag of Words (BoW) and its enhancements [22], [14]. Despite the success of these techniques, CNN-based strategies tend to surpass them in performance. Within this domain, there are two primary strategies: one leveraging classification-trained features for retrieval tasks [1], and another applying deep metric learning for end-to-end training of retrieval systems [9], [28]. The former focuses on strategically using selective search and weighting to apply features optimally, whereas the latter concentrates on refining the metric used for measuring similarities.

3 Methodology

This paper proposes a dual-network approach for image retrieval tasks incorporating compositional information. Comprising two unique network structures, the compositional and retrieval branches are trained using distinct datasets. In this section, we elaborate on the datasets used, describe the network designs, and provide insights into the implementation process.

3.1 Dataset

For the compositional branch of our method, a compositional classification dataset suffices for training. However, the situation is more complex for the retrieval task dataset. Identifying datasets with similar and dissimilar content and composition content is essential to validate historical image retrieval techniques based on content and composition information.

3.1.1 Composition Dataset

The image composition dataset we use is the KU-PCP dataset introduced by J.-T. Lee et al. [16]. It is a photography composition dataset categorized by human annotation into nine labels: rule-of-thirds, center, horizontal, symmetric, diagonal, curved, vertical, triangle, and repeated pattern. The dataset comprises 4,251 outdoor photos, with 3,169 for training and 1,082 for testing, including 20% of the data having multiple labels. The distribution of images across different categories in the dataset is not entirely balanced. For instance, the rule-of-thirds accounts for 22.6% in the training set and 9.4% in the test set. At the same time, the curved category comprises 5.9% and 6.9% of the training and testing sets, respectively. This disparity stems from the dataset creators' methodology, which applies different techniques across categories. For categories with varied styles, such as the rule of thirds, the dataset creators employed CNN-based methods, necessitating a larger sample size and increasing the number of samples for these categories within the dataset. Conversely, handcrafted detectors were used for some categories, resulting in fewer samples. This approach may pose challenges for reproducing or matching their benchmarks with different models and code implementations. In our study, we utilize a unified model to extract composition information. However, the skewed distribution affects our model's performance, making it less effective in some categories than others.

3.1.2 Retrieval Dataset

This paper presents an efficient approach for retrieval task data collection, assuming that images from adjacent frames within a shot share close content and compositional information, whereas images from different shots differ significantly. A "shot" is the primary film production unit in cinematography, encompassing a sequence of continuous frames. It is more extensive than a single frame but smaller than an entire scene. Shots within the same scene generally share the same theme, although camera settings may vary [12]. For the process from historical video data to selecting anchor, positive, and negative samples, see Figure 1. During the training and evaluating phase, an anchor image is selected from the dataset, and then the subsequent image within the same shot group is a positive sample. Conversely, a random frame from another shot is chosen as a negative sample. Despite negative samples and anchor images potentially sharing the same scene, their differing shot groups ensure notable variance in content and composition. Thus, employing shot boundary information allows for extracting image groups from films that are similar and dissimilar in content and composition. These groups can then be further refined by excluding specific shot types.

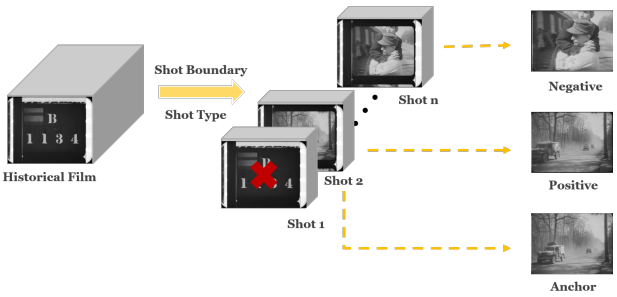

Figure 1: Illustration of converting historical films into pairs of images for retrieval experiments.

We utilized the HISTORIAN dataset proposed by D. Helm et al. [11], encompassing 98 films annotated with 10,593 shots, complete with shot type and boundary details. From this comprehensive collection, images were strategically selected by capturing the first and last frames of every ten frames within each shot, with a maximum of seven images per shot. After excluding Intertitle (I) and Not Available/None (NA) shot types, our refined dataset consisted of 8,432 images from 33 films (6,778 for training and 1,654 for testing). These images span shot types such as Extreme Long Shot (ELS), Long Shot (LS), Medium Shot (MS), and Close Up (CU), reflecting the theme of liberating Nazi concentration camps during World War II. The dataset covers diverse categories, including portraits, architecture, vehicles, and natural landscapes, categorized by various compositional rules. Furthermore, images were cropped according to over-scanned area data to refine the focus on relevant visual information.

3.2 Network Architecture

This section details the architecture of our proposed Composition Clues Network (CCNet) and Content-Based Image Retrieval Network (CBIRNet), explaining how they enhance historical image retrieval by considering both composition and content information. The overall architecture is illustrated in Figure 2.

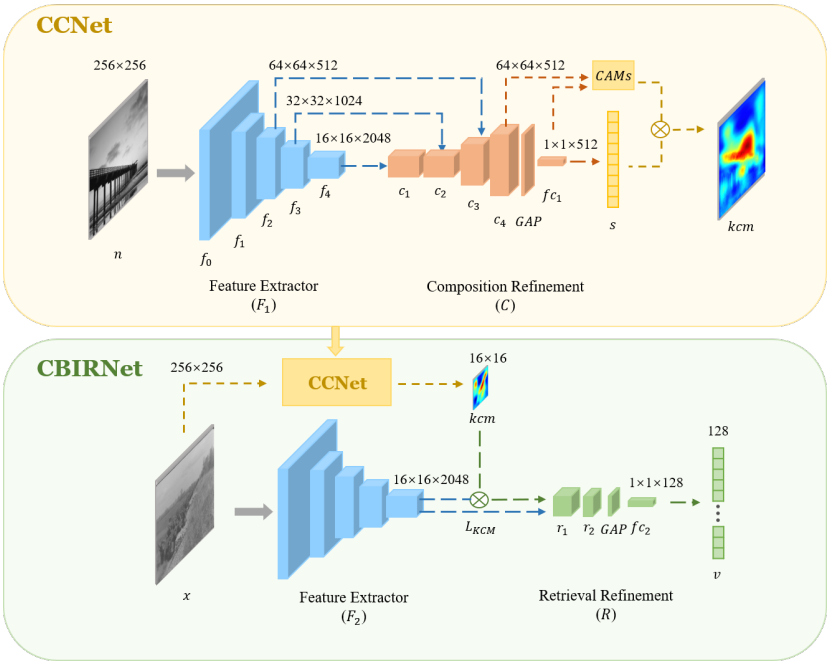

Figure 2: Illustration of our proposed CCNet and CBIRNet.

CCNet is designed to extract composition-related cues from images, operating as a supervised classification model that categorizes composition types. It employs a pre-trained ResNet50 as its backbone architecture (F_1), enabling feature extraction from single-channel input images. F_1 consists of a preprocessing step (f_0) and four main components (f_1 to f_4), extracting progressively deeper features from grayscale images during training and testing stages. Features extracted by the f_4 component of F_1 are fed into a composition feature refining network (C), structured around four CNN layers (c_1 to c_4). Notably, features from F_1's f_2 and f_3 layers are integrated into the c_3 and c_2 layers of C, respectively. This integration strategy enables C to capture composition features across multiple scales effectively.

The features extracted by C are aggregated and then classified through Global Average Pooling (GAP) and a fully connected layer (fc_1) to predict the composition type of the image. Additionally, CCNet employs the KCM mechanism from CACNet [13], utilizing weighted aggregation of $CAMs$ from different composition types to provide an intuitive interpretation of image composition. Each CAM specifically targets an activation map of a given category, with the category's predicted score mapped back to a preceding convolutional layer to generate the CAM. This process effectively highlights the decision-making regions associated with a specific category. By integrating the predicted scores s for all categories with the $CAMs$ through a weighted fusion, we can pinpoint the regions of interest that the model relies on to make decisions in the task of image composition classification. Such integration highlights the areas within the image containing compositional information, resulting in kcm, representing a weighted map of composition information. Using this regional information to guide retrieval allows the retrieval model to focus on locations strongly related to compositional information. Consequently, by assessing whether the compositional areas of two images are consistent, composition is incorporated into the criteria for determining image similarity.

The other component, CBIRNet, consists mainly of a feature extractor F_2 and a retrieval feature refining network R, using the widely adopted ResNet50 as its backbone. Historical grayscale image x is inputted into the network during training and testing. The features $F_2(x)$ extracted from x by F_2 are fused with the kcm obtained from inputting x into CCNet, following a predefined scale L_{KCM}. L_{KCM} is a ratio value ranging from 0 to 1, representing the proportion of kcm's influence in the final fused feature map. For example, when L_{KCM} is set to 0.8, the final feature map is obtained by taking 80% of the input feature map multiplied by L_{KCM} and adding it to 20% of the input feature map. This process yields a feature map that emphasizes areas relevant to compositional cues, which is then dimensionally reduced through subsequent layers r_1 and r_2. Through a series of convolutions, global pooling (GAP), and a final fully connected layer (fc_2), the network transforms image features into a relatively compact feature vector v that retains both content and composition information for subsequent image retrieval tasks.

3.3 Implementation Details

Our model is implemented within the PyTorch framework. Throughout the training process for both tasks, images are resized to 256 × 256 and normalized within a range from 0 to 1. For optimization, we employ the Adam optimizer, initializing the parameters of all networks using the Xavier method [8] and loading pre-trained model parameters for the backbone. CCNet and CBIRNet utilize distinct loss functions to optimize performance. CCNet employs a cross-entropy function to compute losses across various categories. Conversely, the training of CBIRNet employs a cosine embedding loss strategy, which calculates the losses between the anchor image and positive samples, as well as between the anchor image and negative samples, to balance the similarities and differences among images.

4 Experiments

Our experiment investigated whether the compositional information from the CCNet composition model ultimately aids the CBIRNet retrieval model in achieving better image retrieval results. The specific approach involved adjusting the influence parameter L_{KCM} within CBIRNet to examine its impact. We evaluated the model's performance using multiple metrics, including custom-designed ones, to conduct a comparative assessment.

4.1 Image Composition

For the image composition network, we evaluated our model on the grayscale KU-PCP dataset. Our model achieved an accuracy of 0.73, precision of 0.71, recall of 0.70, and an F_1 score of 0.70. Considering our task involves removing crucial color information, the significant reduction in information makes composition classification more challenging. We especially, given the dataset's unbalanced distribution and the high variability of samples, make it difficult for a relatively simple model to focus on compositional information. Figure 3 showcases the impact of our CCNet-trained model through visualizations of KCM. The model's ability to identify critical compositional areas across various compositional rules underscores its effectiveness. Given the accuracy and other metrics of CCNet, we are confident that it has significantly learned to capture compositional rules, making it a valuable tool for training subsequent CBIRNet models.

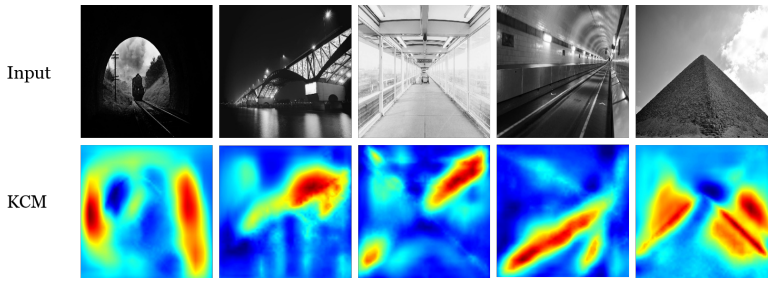

Figure 3: Visualisation of KCM effect. The top row features the original grayscale images, and the bottom row highlights the KCM, pinpointing key compositional areas as detected by our model.

4.2 Image Retrieval

For CBIRNet, we fix the parameters of the CCNet obtained from previous training. By computing the cosine embedding loss between positive samples and the anchor and between negative samples and the anchor, we aim to reduce the distance between positive samples and the anchor while maximizing the distance between negative samples and the anchor.

4.2.1 Quantitative Analysis

Two metrics are employed to assess model performance. The first metric is similar to the one used during training, where we calculate the cosine embedding loss between an anchor image with a positive sample and a negative sample to compare which model has the most negligible loss. The second metric calculates the cosine similarity between the anchor and two positive samples and between the anchor and two randomly chosen negative samples. Ideally, the cosine similarity between the anchor and positive samples should be higher than between the anchor and negative samples. When L_{KCM} equals 0.5, the similarity distribution of all validation set samples with their corresponding four samples is shown in Figure 4. This distribution suggests our model possesses a certain degree of discriminative ability, capable of generating varied scores based on the distinct characteristics of the samples.

A comparison of model metrics trained with varying L_{KCM} values is shown in Table 1. Besides calculating the average positive sample-to-anchor similarity and the average negative sample-to-anchor similarity, we propose a straightforward method of comparing these positive similarities against the negative ones. The model earns a point for each case where the similarity of a positive sample to the anchor surpasses that of a negative sample to the same anchor. The model's average similarity score is then determined by averaging these points across all anchor samples. In an ideal scenario, the optimal model would attain a score of 4, where scores nearing 4 indicate exceptional performance. This method provides a nuanced and equitable assessment of model efficacy. Models without added compositional information, except for the average similarity of negative samples, performed worse on other metrics than those augmented with KCM.

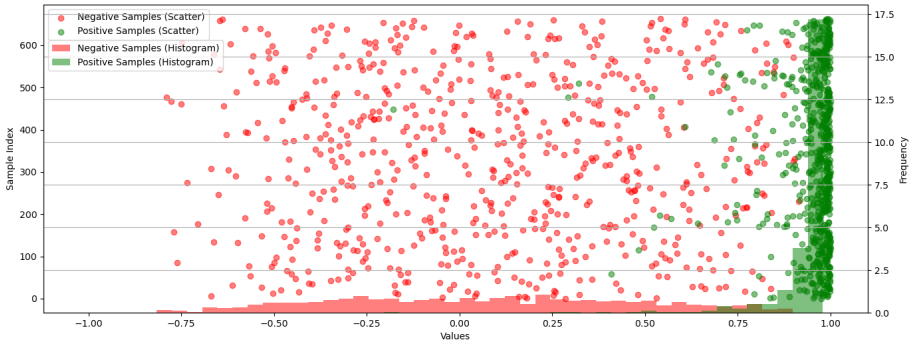

Figure 4: Scatter plot and histogram of positive and negative samples when L_{KCM} is 0.5.

Table 1: Effect of L_{KCM} Values on Model Metrics

L_{KCM}	Cosine Embedding Loss	Avg Pos Similarity	Avg Neg Similarity	Score
0	0.2438	0.9423	0.0277	3.9759
0.5	0.2518	0.9561	0.0864	3.9819
0.8	0.2426	0.9480	0.0378	3.9820

4.2.2 Qualitative Analysis

As shown in Figure 5, we present two sets of specific retrieval examples. Although all three models could identify the target image within the test sample set, indicating their ability to retrieve based on content, the model incorporating compositional information displayed a clear advantage when the alternative images came from different shots with slight variations.

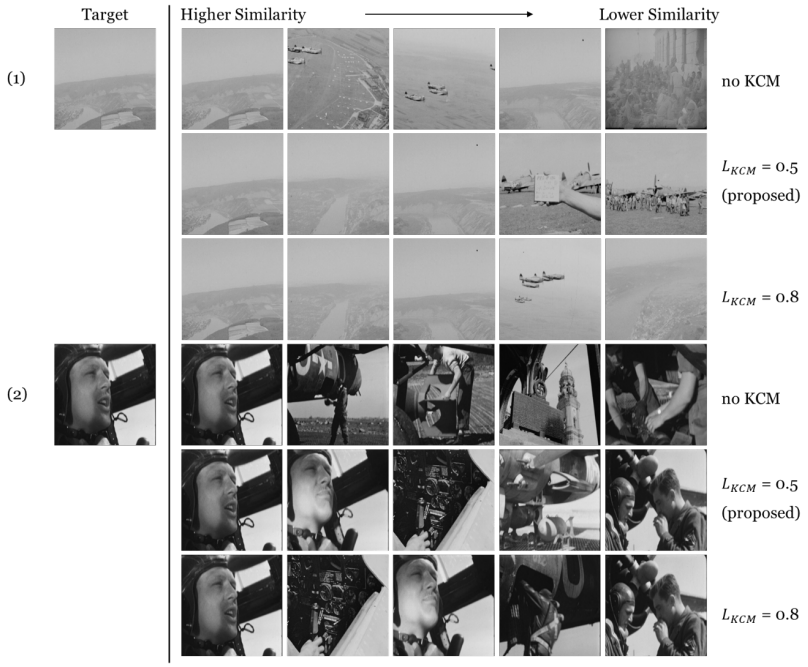

Figure 5: Comparison of retrieval results with different L_{KCM}. We selected only the central frame image from each shot in the test set as the target database for retrieval, returning the five highest similarity-scored images for a single image query.

In the first set of tests, models without compositional information, as shown in the first row, could identify some content-similar images, but it does not keep the composition consistent. The second and third rows integrated compositional information and consistently maintained the composition style vertically. In the second set of test samples, with L_{KCM} set to 0.5, the model considered the compositional information's suggestion of areas of interest, focusing near the left-side facial area. This scale allowed the model to identify the most similar images that models without compositional information could not, placing them at the highest position after the original image. However, the third row, with a higher proportion of compositional information, resulted in insufficient content information, leading to worse outcomes.

Integrating qualitative case studies with quantitative metric analyses demonstrates the superior performance of our model when utilizing an L_{KCM} value of 0.5. Our approach effectively harnesses compositional elements and semantic content, yielding more precise and contextually relevant search outcomes in historical image databases. These findings underscore compositional information's crucial role and promising potential in this domain.

5 Conclusion and Future Work

This paper introduces an image retrieval method that integrates compositional cues and semantic information, utilizing a training and testing pipeline designed around historical images. Our compositional cues network and retrieval network were trained on the KU-PCP composition dataset and the HISTORIAN historical video dataset, respectively, and jointly tested to evaluate the impact of compositional information on retrieval outcomes. We transformed the training of our composition feature extractor into an image composition classification task. Despite needing to convert the KU-PCP Composition dataset into grayscale images to fit our historical image task and the dataset's class imbalance, our image composition classification model still achieved an accuracy of 0.73. This accuracy confirms its effectiveness in extracting compositional information from grayscale images. In retrieval tasks, models incorporating compositional information outperformed those relying solely on content-based methods, both quantitatively and qualitatively. Based on the distances between the features of positive and negative samples, our evaluation metrics suggest that models integrating compositional information have significant potential.

However, to further validate the effectiveness and rationale of our approach, we need to design more experiments and optimize our current model from multiple perspectives. Specifically, to prove that the model can indeed identify images that are closer in both composition and content, constructing a dataset with image pairs that are either similar or dissimilar in composition and content is essential. In this paper, we employed a convenient method of extracting frames from the same shot or different shots within the historical video dataset to construct image pairs. While this method efficiently generates a large amount of training data, it lacks specific important pairs, such as those with similar composition but dissimilar content, and vice versa, since the way we generate image pairs binds the similarity of composition and content. Similarly, since we are working on content-based image retrieval, our dataset also needs pairs where the composition and content category are similar, but the content entities are not. Therefore, although our data processing method is effective, the contribution of building a dedicated dataset is irreplaceable and can better evaluate our model from all aspects.

Regarding the model itself, we proposed a dual-network approach to extract both compositional and content information. Merging these pieces of information has become one of the critical challenges. In our work, direct fusion has proven beneficial for the outcome. However, a more sophisticated fusion method could convincingly allow compositional information to play as significant a role as possible. Thus, our future work on this topic will explore these directions, seeking better ways to extract different types of information from historical images.

Acknowledgments and Disclosure of Funding

This work was supported by the Austrian Science Fund (FWF) – doc.funds.connect, under project grant no. DFH 37-N: "Visual Heritage: Visual Analytics and Computer Vision Meet Cultural Heritage.".

References

[1] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. "Neural codes for image retrieval". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer. 2014, pp. 584–599.

[2] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. "A framework for photo-quality assessment and enhancement based on visual aesthetics". In: *Proceedings of the 18th ACM international conference on Multimedia*. 2010, pp. 271–280.

[3] Wei Chen, Yu Liu, Weiping Wang, Erwin M. Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S. Lew. "Deep Learning for Instance Retrieval: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.6 (2023), pp. 7270–7292.

[4] Gregory Crane and Clifford Wulfman. "Towards a cultural heritage digital library". In: *2003 Joint Conference on Digital Libraries, 2003. Proceedings.* IEEE. 2003, pp. 75–86.

[5] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. "Image retrieval: Ideas, influences, and trends of the new age". In: *ACM Computing Surveys (Csur)* 40.2 (2008), pp. 1–60.

[6] Soma Debnath and Suvamoy Changder. "Computational approaches to aesthetic quality assessment of digital photographs: state of the art and future research directives". In: *Pattern Recognition and Image Analysis* 30 (2020), pp. 593–606.

[7] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. "Multi-modal transformer for video retrieval". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer. 2020, pp. 214–229.

[8] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 249–256.

[9] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. "End-to-end learning of deep visual representations for image retrieval". In: *International Journal of Computer Vision* 124.2 (2017), pp. 237–254.

[10] Mai Lan Ha, Vlad Hosu, and Volker Blanz. "Color composition similarity and its application in fine-grained similarity". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020, pp. 2559–2568.

[11] Daniel Helm, Fabian Jogl, and Martin Kampel. "Historian: A Large-Scale Historical Film Dataset with Cinematographic Annotation". In: *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2022, pp. 2087–2091.

[12] Daniel Helm, Florian Kleber, and Martin Kampel. "HistShot: A Shot Type Dataset based on Historical Documentation during WWII." In: *ICPRAM*. 2022, pp. 636–643.

[13] Chaoyi Hong, Shuaiyuan Du, Ke Xian, Hao Lu, Zhiguo Cao, and Weicai Zhong. "Composing photos like a photographer". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 7057–7066.

[14] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. "Aggregating local descriptors into a compact image representation". In: *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE. 2010, pp. 3304–3311.

[15] Florian Kleber, Stefan Fiel, Markus Diem, and Robert Sablatnig. "Cvl-database: An off-line database for writer retrieval, writer identification and word spotting". In: *2013 12th international conference on document analysis and recognition*. IEEE. 2013, pp. 560–564.

[16] Jun-Tae Lee, Han-Ul Kim, Chul Lee, and Chang-Su Kim. "Photographic composition classification and dominant geometric element detection for outdoor scenes". In: *Journal of Visual Communication and Image Representation* 55 (2018), pp. 91–105.

[17] Xiaoqing Li, Jiansheng Yang, and Jinwen Ma. "Recent developments of content-based image retrieval (CBIR)". In: *Neurocomputing* 452 (2021), pp. 675–689.

[18] Yi Li, LO Shapiro, and Jeff A Bilmes. "A generative/discriminative learning algorithm for image classification". In: *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. Vol. 2. IEEE. 2005, pp. 1605–1612.

[19] Dong Liu, Rohit Puri, Nagendra Kamath, and Subhabrata Bhattacharya. "Composition-aware image aesthetics assessment". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2020, pp. 3569–3578.

[20] Ligang Liu, Renjie Chen, Lior Wolf, and Daniel Cohen-Or. "Optimizing photo composition". In: *Computer graphics forum*. Vol. 29. 2. Wiley Online Library. 2010, pp. 469–478.

[21] Markus Mühling, Manja Meister, Nikolaus Korfhage, Jörg Wehling, Angelika Hörth, Ralph Ewerth, and Bernd Freisleben. "Content-based video retrieval in historical collections of the German broadcasting archive". In: *International Journal on Digital Libraries* 20 (2019), pp. 167–183.

[22] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. “Large-scale image retrieval with compressed fisher vectors”. In: *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE. 2010, pp. 3384–3391.

[23] Leo Sampaio Ferraz Ribeiro, Tu Bui, John Collomosse, and Moacir Ponti. “Scene designer: compositional sketch-based image retrieval with contrastive learning and an auxiliary synthesis task”. In: *Multimedia tools and applications* 82.24 (2023), pp. 38117–38139.

[24] Donghee Sinn. “Impact of digital archival collections on historical research”. In: *Journal of the American Society for Information Science and Technology* 63.8 (2012), pp. 1521–1537.

[25] Xiaoou Tang, Wei Luo, and Xiaogang Wang. “Content-based photo quality assessment”. In: *IEEE Transactions on Multimedia* 15.8 (2013), pp. 1930–1943.

[26] Jingdong Wang and Xian-Sheng Hua. “Interactive image search by color map”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 3.1 (2011), pp. 1–23.

[27] Xiang-Yang Wang, Bei-Bei Zhang, and Hong-Ying Yang. “Content-based image retrieval by integrating color and texture features”. In: *Multimedia tools and applications* 68.3 (2014), pp. 545–569.

[28] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. “Multi-similarity loss with general pair weighting for deep metric learning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 5022–5030.

[29] Hongyang Yan, Mengqi Chen, Li Hu, and Chunfu Jia. “Secure video retrieval using image query on an untrusted cloud”. In: *Applied Soft Computing* 97 (2020), p. 106782.

[30] Shiliang Zhang, Qi Tian, Ke Lu, Qingming Huang, and Wen Gao. “Edge-SIFT: Discriminative binary descriptor for scalable partial-duplicate mobile search”. In: *IEEE Transactions on Image Processing* 22.7 (2013), pp. 2889–2902.

[31] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. “Learning deep features for discriminative localization”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2921–2929.

[32] Wengang Zhou, Houqiang Li, Richang Hong, Yijuan Lu, and Qi Tian. “BSIFT: Toward data-independent codebook for large scale image search”. In: *IEEE Transactions on Image Processing* 24.3 (2015), pp. 967–979.