Proceedings of 10th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2021 Month: October Volume: 10

Content Based Image Retrieval Using Convolutional Neural Network, Principal Component Analysis and K-means Clustering

Prashant Adhikari [a], Anish Pandey [b], Sanjeeb Prasad Panday [c]

[a, c] *Department of Electronics and Computer Engineering, IOE, Tribhuvan University, Nepal*
[b] *Electronics and Communication Engineer, National Institute of Technology Durgapur, India*
Corresponding Email: [a] 075msice015@pcampus.edu.np, [b] sharmaanix@gmail.com, [c] sanjeeb@ioe.edu.np

Abstract
Content based image retrieval (CBIR) is a system that takes an image as an input and provides a set of similar images to the input as output in an order of matched similarity. Features matching among images is a vague topic. It depends on what characteristics are taken into account and what degree of characteristics are balanced. In terms of semantics, images with high feature similarities to the query can be quite distinct from the query. This paper introduces a scheme for image retrieval, cluster-based image retrieval via unsupervised learning. In a certain feature space, semantically similar images appear to be clustered. This paper aims to capture semantic concepts by comparing images with the same semantics, and extracting image clusters will provide more reliable results instead of a collection of ordered images. Clusters therefore offer the algorithm and the users semantic relevant clues obtained from the input question that indicate where to navigate. Principal component analysis is used to reduce the dimensions of the extracted features of the training images. Here a representative image that has feature similarity to majority of the images, also known as cluster centre is compared with the input query processed result and then images of the cluster are sorted as per the similarity measure in terms of feature. Oxford university 17 category flower data set is used in this paper. Precision parameter is calculated based on the number of similar image retrieval out of five most similar images and the system is evaluated using the average precision for different test cases.

Keywords
Feature Extraction, Principal component analysis, K-means clustering

1. Introduction

Rapid growth of internet, social media and image posting behavior of internet users have increased the image repositories in a larger way. Effective image retrieval from a wide archive of images is still a difficult task. Another ambiguous job in the modern digital world is identifying similar features in images and receiving those images. Nowadays, images are stored in image repositories via use of CBIR approach according to the visual details of the query image. CBIR is considered to be an accessible and advanced area of research that provides solutions for different problems presented by image similarity analysis. CBIR methods use local characteristics such as form, texture, color and spatial image structure for image retrieval. Semantic gap means how the machine stores the images at pixel level and how the human as a whole perceives the data. Several works were conducted to resolve this level of semantic gap. Visual similarity in CBIR is the baseline approach. The similarity between function vector values belonging to various semantic categories decreases CBIR's output since images without a semantic connection are retrieved. A feature descriptor defines an image in such a way that the visual content extent of similarity can be calculated to calculate the visual similarity between the images. In addition, the query image feature vector and equated image feature vectors are compared and similarity index is determined, which defines the images retrieved. Image feature descriptors are commonly known as descriptors of local or global features. As an alternative to retrieving a set of ordered images, a cluster based image retrieval scheme may be used. In addition to the feature similarity of the images to the query, the image clusters are obtained from an unsupervised learning process. But also how pictures are equivalent

to one another. It attempts to rely on the indication that images with the same semantics are identical. It is a similarity-driven approach that can be based on practically any symmetrical calculation of real-value image similarity.

Different techniques have been established for content based image retrieval in the last three decades. Several features can be considered for extraction such as color, texture, boundary, contour and others for feature extraction. Local characteristics of an image can be considered for some transformations and illumination. Different techniques have been practiced for feature extraction such as BRISK, SIFT, SURF, rotated BRIEF and others. Some methods have also used Visual Bag of Words. Fusion of different techniques have also been practiced by some researchers. Features computed using combination of different techniques form a high dimensional feature descriptors. K-means clustering is applied to these integrated features and presented in a visual space. Some researchers have also initiated a coherent approach that are based on evolutionary computation theory. Evolutionary model based on differential evolution have also been chosen by some authors. Some crossover method have also been applied to increase the efficiency of differential evolution. Particle Swarm Optimization technique has also been developed which facilitated the combination of multiple genetic algorithms. A paper published in 2013 propose an improvisation over existing CBIR technique. In addition to the normal feature descriptors like color and texture, it also used wavelength based histogram approaches. Some researchers have also proposed artificial neural networks in which feature fusion technique can be implemented. During the application of artificial neural network, dimensionality reduction technique is applied to reduce the complexity of problem. Features are balanced for determining the similarity between input and stored image. In another research published in 2014, different techniques were used to form the feature vectors like edge histogram, color layout, scalable color. The whole principle is based on Visual Bag of Words. Presence of interference regions in images is sorted out with the use of semantically significant region and region significant index. For images with noisy inclusions like tags, a researcher proposed in 2017 that semantic space is built on the both visual and textual properties and when labelled denoising is performed, efficiency of such image retrieval is noteworthy. Recent paper published in 2019 has suggested the use of Visual Bag of Words layout that uses the principle of both SIFT and BRISK. This approach describes image features as clusters. Both BRISK and SIFT generate separate dictionaries and both of them are combined. Histograms are constructed on the basis of combined dictionary constructed. The working principle is as follows: The image repository is divided into two sets of data: research and training set. Features are extracted using both SIFT and BRISK. Individually extracted features are then submitted to a K-means clustering algorithm. A combined dictionary is constructed using both dictionaries. Visual words are then constructed from a combined result of both and histograms are constructed from it. Classifier is learned and similarity is measure between the input image and the images in repository.

Such tests are conducted in the different image repositories. When individually applying SIFT and BRISK, the efficiency of the system is compared and it is found that the combination of both SIFT and BRISK offers excellent system performance than individually applying SIFT or BRISK.

In this regard, this paper proposes a CBIR system based on clustering approach with reduced dimensions of features. In previous papers, clustering techniques have been applied to the whole features extracted. This paper introduces a concept of reducing dimension before performing clustering on the extracted features. Principal component analysis has a significant role in covering the major features of the objects and reduce the time of image retrieval with significant performance of CBIR system. It generalizes the features extracted from the dataset, minimizes the dimensions and apply a cluster based scheme to those features. It aims to capture the semantics in those images and provides most similar images to the query images in more efficient manner. Choice of ResNet for the feature extraction is based on the increased training parameters of ResNet over others. VGG-16 has roughly 138 million parameters while ResNet has 25.5 million parameters. ResNet employs the use of thinner and deeper networks. Choice of 1*1 convolutions before applying 3*3 convolutions helps to reduce number of operations. ResNet have different variants like: ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-110 and others. Out of these, ResNet-50 is used in this thesis considering computational effectiveness and considerable accuracy in feature representation.

K-means clustering has been chosen for clustering because of its simplicity and applicability for large datasets.

2. Methodology

2.1 Proposed System

Figure 1 shows the proposed methodology of the cluster based CBIR system. It consists of data pre-processing unit, feature extraction unit, PCA unit for reducing dimension and clustering unit. Detail description of methodology is as below:

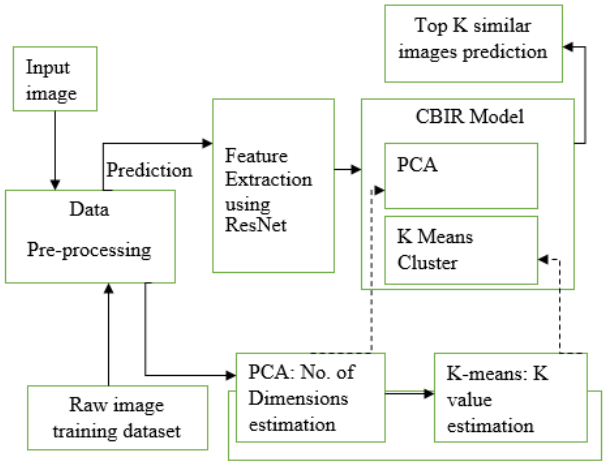

Figure 1: System methodology of cluster based CBIR system

Selection Of Dataset Dataset selection process is the first step of any content based image retrieval paper. Following things should be considered while selecting a proper dataset:
i) Amount of data should be sufficient to train the model
ii) Data should not be out of scope. Even if the data set is diverse, each type of data should contain multiple similar data.
iii) Ultimate goal of data set is to form a good cluster. Dataset is selected from flower dataset of Oxford university research group that contains about 1500 images of different types and categories.

Preprocessing of image dataset Images available in the data set can be of different shapes and sizes. Therefore it is required to resize the images into a uniform sizes. Also, it is required to normalize the image. Normalization brings the pixel intensity to 0 to 1. This makes the computation efficient. Also, normalization can avoid the influence of high frequency noise and very low noise. There are several method of normalization like using NORM_MINMAX or NORM_L1. When using a NORM_L1, each pixel value is split by the sum of the absolute values of all the image pixels. It normalizes in such a way that the minimum destination value is alpha defined when using a NORM_MINMAX, and the maximum destination value is beta specified. NORM_MINMAX uses only scales and shifts. It adds constants and multiplies by constants. Resizing at a size of 224 by 224 pixels and normalization using NORM_L1 has been done in this paper.

Feature Extraction The more features we have, the more data we need to train a good model. For a limited number of training data, model's accuracy will decrease for every feature we have. Features can be number of pixels for an image. For a 64*64 images, we have 4096 features. Feature has been extracted using ResNet-50 model. Weights are pre-specified in ResNet-50 and images are convolved through different layers, pooled and finally generated a feature vector. Dimension of single image after feature extraction is found to be 100352.

Dimension Reduction and Optimization of Components Compression could be a solution to reduce the number of features. The ultimate essence of compression is finding a way to keep as much information as possible about the image without losing the essential structure. Dimensionality reduction can actually improve the results of the model. While trying to reduce the dimension, we must find the number of dimension that keeps about ninety percent of the variance of the original image. The principal component analysis is used for dimension reduction. Principal component analysis is a method of decreasing dimensionality that is often used to decrease the dimensionality of large data sets by converting a large data set of variables into a smaller one that still retains much of the large set of information. It is widely used to minimize dimensions by papering each data point on only the first few key components to obtain lower dimensional data while maintaining the variance of the data as much as possible. The main components are essentially the covariance matrix's eigen vectors. The key components are also determined by the data covariance matrix's eigen decomposition or the data matrix's singular value decomposition.

During optimization of components, the feature

values of components is set to thousand. Model parameters are learned in this optimization of components. For example, from a training set, mean and standard deviation for normalization are learned from. Explained variance ratio is calculated. The ratio of variance explained is the proportion of variance explained by each of the components selected. If n components are not specified, all components are stored and the ratio total is equal to one. Here the sum of explained variance ratio is made approximately ninety percent. Dimensions have been reduced from 100352 to 1000 and optimized to 122 dimensions. 122 dimensions gave us a total 90.06 percentage of cumulative explained variance.

K-means Clustering K-means clustering is a type of unsupervised learning used when we have unlabeled data. Unlabeled data means data without classes or categories. The ultimate aim of this algorithm is to find the number of groups in the data, with the variable K representing the number of groups. Based on the features given, it works iteratively to allocate each data point to one of the K groups. Clusters are generated based on the similarity of features.K-means clustering is performed to establish clusters of associated images. Clusters numbers run from minimum one to three hundred. Minimum distortion inertia and maximum distortion inertia is determined if the clusters made are compact or not.

Choosing value of K The clusters and data set labels are found by K-means Clustering for a unique pre-chosen K. The user needs to run the K-means clustering algorithm for a range of K values and compare the results to find the number of clusters in the data. In general, there is no formula for determining the exact value of K, but using the following techniques, a reliable approximation can be obtained.

The mean distance between data points and their centroid cluster is one of the metrics widely used to compare outcomes across various K values. As the distance to data points is always decreased by increasing the number of clusters, increasing K would always reduce this metric to the extreme of reaching zero when number of data points is same as K. Thus, as the sole target, this metric cannot be used. Instead, the mean distance to the centroid is plotted as a function of K and the ”elbow point,” where the rate of reduction changes sharply, can be used to calculate K approximately.Elbow point is obtained near 17 number of categories. So, number of clusters is fixed to 17.

Similarity Matching The input image is first pre-processed and normalized. Then, it is represented by feature vector. After dimension reduction via principal component analysis, it is reduced to reduced feature vector. Then, its match with the cluster is found out to find out the similar images to the input images. Cosine similarity is found out between the input image and the corresponding cluster images to identify the top similar images. Similarity can be observed in the output images in terms of petal and flower structure as represented by the feature vectors. Percentage of match between the input image and the output images is given by the similarity index.

Similarity Ranking Similarity between the feature vectors is calculated using cosine similarity and top k images from the cluster is shown as output. Similarity is observed in terms of structure and shape of petals and flowers in the output images

2.2 Model Description

Data Pre-processing Raw input image is provided to the data processing and feature extraction unit. Pre-processing section includes two main processes: resizing of images and normalization of images. Resizing is done to make the images of same size. Size of 224 by 224 pixels is taken in this paper. Normalization is done to avoid the influence of high frequency and low frequency noise. Normalization brings the pixel intensity from 0 to 1.

Feature Extraction Using ResNet-50 The more features we have, the more data we need to train a good model. For a limited number of training data, model’s accuracy will decrease for every feature we have. Features can be number of pixels for an image. For a 64*64 images, we have 4096 features. Images from repository are passed into the ResNet-50 model and then corresponding feature vectors are extracted. ResNet-50 is a pre-trained model which has pre-designed weight vectors. Images are passed through different layers and different standard operations like convolution, pooling, ReLu function to generate different feature vectors. Features generated from such network are then used for further analysis. ResNet works excellent with classification tasks so this model is used in feature extraction.

CBIR Model CBIR model has two components:

PCA component Principal Component Analysis is a dimension reduction technique. For efficient PCA result, several parameters need to be fixed like: number of components after reduce, transformation operation, explained variance ratio. Generally, number of components needs to be specified in such a way that the explained variance ratio needs to be around ninety percentage.

K-means Cluster K-means clustering is done to make the groups of related images. For a fine clustering result, parameters like number of clusters needs to be fixed to its proper value. The elbow point of the plot must be located between the mean distance between the point and the centroid cluster, where the rate of decrease changes sharply.

Training Process For training process, we need to fix the PCA model and the K-means cluster model. During the training process, values are kept at random and the effect of such values in the result is analyzed. The exact value of the parameters to be tuned is not defined by any unique procedure. Instead, we randomize the parameters to some values and observe its effects accordingly. The best value where the desired effects are observed are taken as tuned value.

Selection of top K similar images After the training process and parameter tuning, we need to find top k similar images. This is made based on the comparison of reduced feature vector representation.

Generation of output images Proper images are selected from the top K similar images and these are represented as output of the system accordingly. Here five most similar images are provided as output from the system.

3. Result Analysis

3.1 Number of Dimensions in PCA

While plotting the cumulative explained variance versus the number of components, the graph obtained is below.The graph is constructed randomly choosing 1000 components and the cumulative explained variance. It shows the cumulative explained variance ratio is nearly 90.06 percentage when the number of components is 122 as shown in Figure 2. Originally feature vectors extracted from ResNet architecture are of components 100352. Such huge dimension is reduced to 122 number of components using principal component analysis which is a great advantage for using the processing power of the machine efficiently.

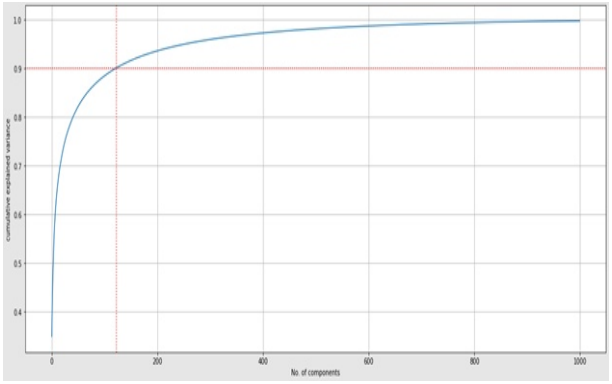

Figure 2: Plot of cumulative explained variance versus the number of components

3.2 Determining the number of clusters in K-means clustering

While plotting the K-means inertia versus the number of clusters, the graph obtained is shown below. To determine the elbow point of the K-means cluster, 300 number of clusters is chosen initially. A graph is plotted between the K-means inertia number of clusters as shown in Figure 3.

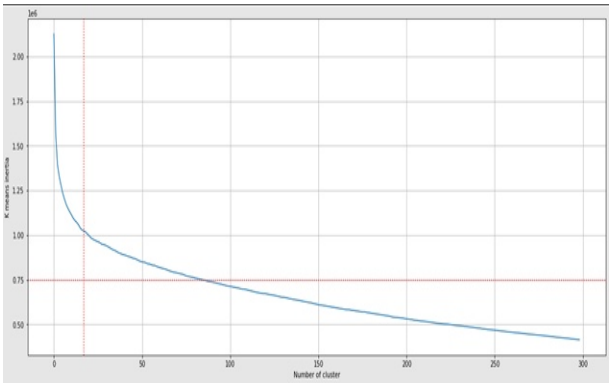

Figure 3: Plot of K-means inertia versus the number of clusters

A point near the sharp decrease of the graph is chosen to be the elbow point. Corresponding cluster axis is obtained to be 17. So, 17 number of clusters are constructed in the data.

3.3 Test Images

Figure 4: Input Lilly Valley Image, Source: Dataset

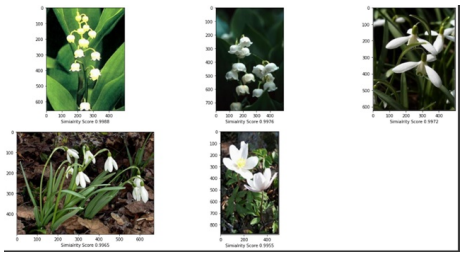

Figure 5: Top 5 similar images for the input Figure 6

Figure 6: Input Sunflower Image, Source: Dataset

Figure 7: Top 5 similar images for the input Figure 8

Figure 8: Input Fritillary Image, Source: Dataset

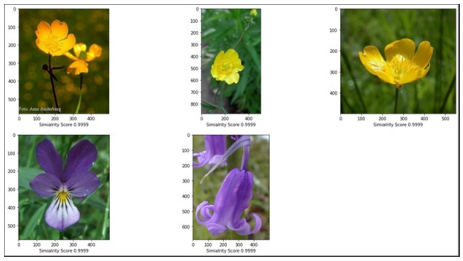

Figure 9: Top 5 similar images for the input Figure 10

Figure 10: Input Rose Image, Source: pexels.com

Figure 11: Top 5 similar images for the input Figure 12

Features extracted from residual convolutional neural network: ResNet-50 have 100352 dimensions. These dimensions are reduced to 122 applying principal component analysis. Reduced features are then clustered in 17 clusters applying K-means clustering algorithm. K-means clustering is chosen because of its low computational complexity. Cluster corresponding to the input image is identified based on the features of input images obtained from ResNet-50 model. Then, the cosine similarity of the input image with the other flowers images present in the cluster is obtained. These cosine similarities are ranked in descending order and the top 5 images are then displayed as output. Test sets are then selected on the basis of flower having different petals length,

Figure 12: Input Image, Source: pexels.com

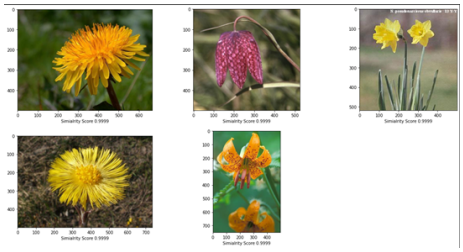

Figure 13: Top 5 similar images for the input Figure 14

Figure 14: Input Image, Source: pexels.com

Figure 15: Top 5 similar images for the input Figure 16

Figure 16: Input Image, Source: pexels.com

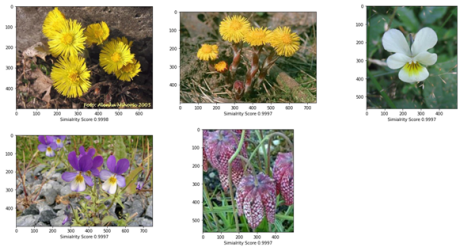

Figure 17: Top 5 similar images for the input Figure 18

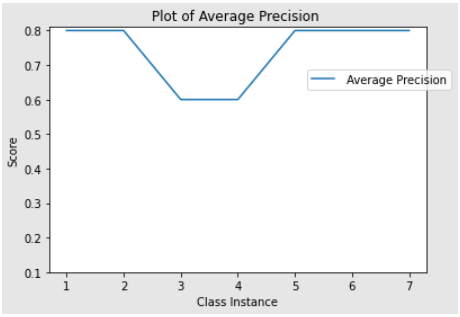

Figure 18: Plot of average precision of test images

structure and number. Test set corresponding to Figure 6, Figure 8 and Figure 10 are taken from the training data set. While evaluating the result, it is observed input image Figure 6 belong to cluster number 8, Figure 8 belong to cluster to cluster number 7 and Figure 10 belongs to cluster number 17. Top 5 similar images obtained for these input images can be observed similar to the input image in terms of petal structure, number of petals and type of flower. Figure 10 does not belong to the training data set. These images are obtained from open source. Categories of flowers corresponding to Figure 10, Figure 12, Figure 14 and Figure 16 and are not present in the image repository. Despite of the absence of those flowers in image repository, the model search for those flowers images which are similar to the flower structure, petal length, petal numbers, and petal structure of the images present in repository. Predicting the similar images for those open source images is a major advantage of this paper. Similar images obtained from the repository for the input image is based on the feature extracted by ResNet-50 model. If different model is used to extract the features, results could be different. Features similarity within the same cluster has been calculated based on cosine similarity. This classification is based on the assumption that the input image lies in that cluster whose center is at the shortest distance than others. Similarity can be seen in the size and structure of petals for qualitative analysis.

70 different test images were separated from the data set. Five most similar images to the query image have been considered for the plot of Figure 18. Images belonging to the same class are taken as true positives and those not belonging are taken as false positives. Based on these data precision is calculated. Precision is averaged over different classes and it is plotted as shown in Figure 18. Average precision obtained in this paper for the Oxford 17 categories flower dataset is 74.29 percentage.

4. Conclusion

Content based image retrieval using principal component analysis and K-means clustering is an image searching application based on similarity. Currently, ResNet-50, principal component analysis and K-means clustering are the basis of this framework. Features are extracted from the images using ResNet-50, feature vectors are reduced to lower dimension applying principal component analysis and clustering is done on these images based on similarity. Average precision is obtained to be 74.29 percentage for the Oxford University 17 category flower dataset. In this paper, only feature similarity comparison has been done. Object recognition is still some important work to do. This application can be extended by means of neural network application and recognizing the objects in the images to specify which object is actually present in the image. Enhancements can also be made in the feature extraction and clustering mechanisms to improve the result.

References

[1] Uzma Sharif, Zahid Mehmood, Toqeer Mahmood, Muhammad Arshad Javid, Amjad Rehman, and Tanzila Saba. Scene analysis and search using local features and support vector machine for effective content-based image retrieval. *Artificial Intelligence Review*, 52(2):901–925, 2019.

[2] Yixin Chen, James Z Wang, and Robert Krovetz. Content-based image retrieval by clustering. In *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 193–200, 2003.

[3] Berthier Ribeiro-Neto and Ricardo Baeza-Yates. Modern information retrieval. *Addison-Wesley*, 4:107–109, 1999.

[4] Kobus Barnard and David Forsyth. Learning the semantics of words and pictures. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 408–415. IEEE, 2001.

[5] Chad Carson, Serge Belongie, Hayit Greenspan, and Jitendra Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on pattern analysis and machine intelligence*, 24(8):1026–1038, 2002.

[6] Avi Arampatzis, Konstantinos Zagoris, and Savvas A Chatzichristofis. Dynamic two-stage image retrieval from large multimedia databases. *Information Processing & Management*, 49(1):274–285, 2013.

[7] Sergei Vassilvitskii and David Arthur. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2006.

[8] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[9] Yixin Chen, James Ze Wang, and Robert Krovetz. Clue: Cluster-based retrieval of images by unsupervised learning. *IEEE transactions on Image Processing*, 14(8):1187–1201, 2005.

[10] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)*, 40(2):1–60, 2008.