

RESEARCH ARTICLE

Image Annotation With YCbCr Color Features Based on Multiple Deep CNN- GLP

MYASAR MUNDHER ADNAN^{1,2}, WALEED HADI MADHLOOM KURDI^{3,4}, SARAH ALOTAIBI⁵, AMJAD REHMAN⁶, (Senior Member, IEEE), SAEED ALI OMER BAHAJ⁷, MOHAMMED HASAN ALI^{8,9}, AND TANZILA SABA⁶, (Senior Member, IEEE)

¹Faculty of Engineering, School of Computing, Universiti Teknologi Malaysia, Skudai, Johar 81310, Malaysia

²College of Technical Engineering, Computer Techniques Engineering, Islamic University, Najaf 54001, Iraq

³Nursing Department, Altoosi University College, Najaf 54001, Iraq

⁴Department of Electrical Engineering, Faculty of Engineering, University of Kufa, Najaf 540011, Iraq

⁵Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia

⁶Artificial Intelligence & Data Analytics Laboratory, CCIS, Prince Sultan University, Riyadh 11586, Saudi Arabia

⁷MIS Department, College of Business Administration, Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia

⁸College of Technical Engineering, Imam Ja'afar Al-Sadiq University, Al-Muthanna 66002, Iraq

⁹College of Computer Science and Mathematics, University of Kufa, Najaf 540011, Iraq

Corresponding author: Saeed Ali Omer Bahaj (saobahaj@gmail.com)

ABSTRACT Digital image collections are becoming increasingly popular due to their ease of use. Still, the need for adequate indexing information makes it difficult for users to find the specific images they need. With the vast number of digital images generated daily, these databases have become enormous, making accurate image retrieval challenging. One of the most challenging tasks in computer vision and multimedia research is image annotation, where keywords are assigned to an image. Unlike humans, computers can measure colors, textures, and shapes of images but fail to interpret them semantically, known as the semantic gap. This makes image annotation complex. For semantic-level concepts generation the raw image pixels provide not enough Unmistakable information. Which mean for of “words” or “sentences” there is no clear definition with the semantics of an image unlike text annotation. Therefore, this study aims to bridge the semantic gap between low-level computer features and human interpretation of images. The proposed enhanced automatic image annotation system maps multiple labels or into single image, providing an in-depth understanding of the visual content’s meaning. This is achieved by combining Convolutional Neural Networks-based multiple features (Y is the green component of the color, Cb and Cr is the blue component and red component called YCbCr color space and Gaussian–Laplacian Pyramid) and neighbors to recall and balance precision. The image annotation (IA) scheme uses a Global Vectors for Word Representation (GloVe) model with CNN-Gaussian–Laplacian Pyramid and learning representation to predict image annotation (IA) accurately. The proposed image annotation (IA) system was execution on three public datasets and showed excellent flexibility of annotation, improved accuracy, and reduced computational costs compared to existing state-of-the-art methods. The image annotation (IA) framework can provide immense benefits in accurately selecting and extracting image features, minimizing computational complexity and facilitating annotation.

INDEX TERMS Features extraction, YCbCr color, digital learning, Gaussian–Laplacian pyramid, image annotation, technological development.

I. INTRODUCTION

The significance and meaning of images in our lives can be explained by the famous Confucian proverb, “A single image

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson¹⁰.

carries the impact of countless words.” Images can communicate intricate ideas and feelings without the need for extensive explanations. Digital images have become an integral part of our lives in both professional and domestic settings, frequently employed in fields such as security systems, advertising, medicine, insurance, finance, and commerce.

In personal life, digital images identify individuals, animals, locations, and occasions such as birthdays, marriages, vacations, accidents, and sporting events. This pervasive use of digital images with billions of images uploaded to specialized websites has led to a rapidly increase in their quantity. For example, according to statistics from early 2017 [1], [2] 136,000 photos were uploaded to Facebook every minute, 95 million photos and videos were posted on Instagram each day, and over 20,000 snapshots were shared every second on Snapchat [3]. These numbers highlight the growing demand for digital images on social media platforms.

With the proliferation of digital images on various online platforms and personal collections, the size of image databases has reached unprecedented levels. These collections' popularity largely depends on how easily internet users access them. However, many of these image databases need more indexing information, making it challenging for users to retrieve the images they need from anywhere and at any time. Currently, users often need help accessing image information by entering search commands. Therefore, an automated system is required that can efficiently analyze the contents of images in a more meaningful way and instantly retrieve the information required by the user. The significant contributions of this article are summarized below.

1) To incorporate information about multiple feature types into a new vector representation of features for the IA system.

2) Enhance the efficacy of image coding and annotation utilizing deep learning multiple Convolutional Neural Networks with Gaussian–Laplacian Pyramid (DL-MCNN- GLP).

3) Incorporate multiple image descriptors into a merging multi-convolutional neural network model to enhance the image annotation framework's performance and evaluate the proposed scheme's image retrieval system. The next section provides context for prior research reported in the literature. In Section III, a sophisticated technique for the capture is described for subsurface features. The main emphasis of the paper lies in Section IV, where it presents an innovative approach for labeling images. The evaluated of new method in comparison to SEM [3], 2PKNN [4], and CNN-THOP [5]. The conclusion concludes by summarizing the current identifying and findings possibly future research directions.

II. IMAGE ANNOTATION RELATED WORKS

Visual content repositories like image and video-sharing websites need indexing, multimedia information search, and image search. Indexing and querying big image sets is straightforward using annotations. Look at prior research on image annotation and feature extraction.

A. IMAGE ANNOTATION FEATURES

It's shown that image annotation features like colors, textures, structures, and forms can represent all extracted areas. This work characterized each image region with varied features to reduction the computational cost for identifying from the training and testing data the most of the appropriate features

extracted and improve the image annotation algorithm's shape extraction.

1) RED, GREEN, AND BLUE (RGB) COLOR HISTOGRAM

First, each region's R, G and B color channels were quantized into 64 bins and represented by a 64-dimensional histogram. Next, each color histogram feature was normalized such that its components sum is equal to 1 [6].

2) HSV COLOR MOMENTS

Each region was mapped to the HSV color space followed by the computation of the mean, standard deviation, and skewness of the H, S, and V components which were used as features. Finally, this feature subset was represented by a 9-dimensional vector [7].

3) UV COLOR MOMENTS

First, each region was mapped into the LUV color space. Next, the mean, standard deviation, and skewness of the L, U, and V components were computed. A 9-dimensional vector represented this feature subset. The color moments feature subsets of both HSV and LUV were normalized to achieve zero mean and unit standard deviation [8].

4) EDGE HISTOGRAM

A MPEG-7 edge histogram descriptor (EHD) variation was used to express edge frequency and directionality in each image area. A simple edge detector operator identified edges and divided them into five categories: vertical, horizontal, diagonal, anti-diagonal, and non-edge. The EHD has five bins based on category frequencies [9].

5) WAVELET TEXTURE FEATURES

Frequencies and resolutions were evaluated for each area. The Haar filter bank split the image into three scales, yielding 10 components. They have three-scale horizontal, vertical, and diagonal components. The features vector became 20-dimensional after computing each component's mean and standard deviation. Normalizing this feature subset yielded zero mean and unit standard deviation [10].

6) SHAPE FEATURES

Region eccentricity, direction, size, solidity, and extension were calculated. An ellipse using the region's second moments was used to calculate eccentricity. Calculate the ellipse's foci-major axis length next. The orientation was the angle between the ellipse's x-axis and major axis. The region's pixels also specified its area. The region's convex hull pixel proportion determined solidity. The extent was the ratio of pixels in the enclosing box of the regions to its area [11].

7) THE YCBCR COLOR FEATURES

The YCbCr color model is a way of representing colors in an image using three components: Y (luminance), Cb (blue

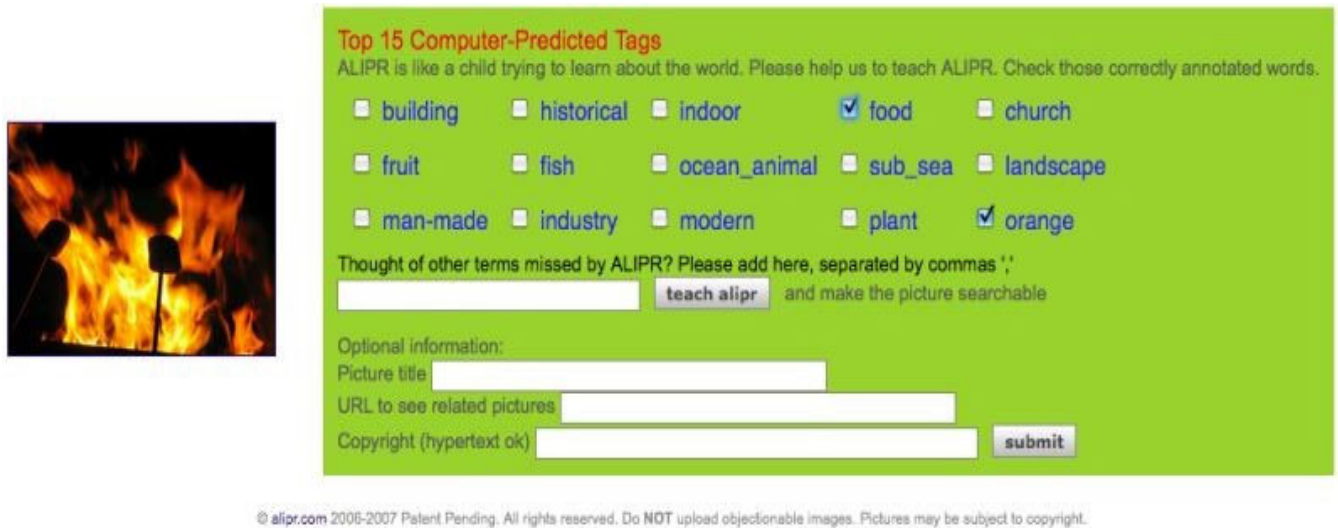


FIGURE 1. Example of annotation of an ALIPR image [15].

chrominance), and Cr (red chrominance). These components are calculated from the original RGB (Red, Green, and Blue) color channels in the following way: Y (Luminance): The Y component represents the brightness or intensity of the color. It is calculated as a weighted sum of the RGB channels [11].

$$Y = 0.299 * R + 0.587 * G + 0.114 * B \quad (1)$$

These coefficients (0.299, 0.587, and 0.114) represent the perceived intensity of the colors by the human eye. Cb (Blue Chrominance): The Cb component represents the blue-difference information, or how much blue differs from a neutral gray color [11]. It is calculated as follows:

$$Cb = (B - Y) * 0.564 + 128 \quad (2)$$

The 0.564 is a scaling factor, and 128 is an offset to center the values on 128. Cr (Red Chrominance): The Cr component represents the red-difference information, or how much red differs from a neutral gray color [11]. It is calculated as follows:

$$Cr = (R - Y) * 0.713 + 128 \quad (3)$$

Similar to Cb, 0.713 is a scaling factor, and 128 is an offset.

B. IMAGE ANNOTATION METHODS

Annotation of images by keywords is a way to associate “semantics” with an image. Each image is assigned a keyword or set of keywords that describe the semantic content of the image. In addition to manual annotation, there are semi-automatic annotations and automatic annotations.

1) MANUAL ANNOTATION

Users manually annotate images with keywords. Ground truths need image bases. These verify automated annotation techniques. Manual annotation is time-consuming and costly, but it better describes an image’s content. Manual annotations

are subjective, depending on the individual annotating the image [12]. States that the same image may have several interpretations. The next experiment proved his theory.

2) AUTOMATIC ANNOTATION

Automated image annotation is the process of automatically assigning descriptive terms from a dictionary to images. The input image is referred to as the target image, while the annotated image is regarded as the Output image. By analyzing color, texture, and shape, computers can calculate low-level features, but they do not comprehend them like humans do. To achieve automatic image annotation, there must be a bridge between the low-level features computed by computers and human interpretations at a semantic level [13]. Researchers have been continuously working on developing automated image annotation methods in recent years. Automatic annotation uses generative, discriminating, and graphic models.

3) SEMI-AUTOMATIC ANNOTATION

Semi-automated annotation bridges manual and automatic annotation. Annotating photos or refining automated annotations involves human input. Image search uses it. The recovery system displays picture results after a keyword search. The user chooses the most relevant input. Images that the user likes will automatically have the query term annotated. Several systems allow picture annotation. ALIPR [14] is an example. Keywords appear when images are uploaded. The user selects keywords from this list or adds keywords related to the picture. Figure 1 illustrates. Semi-automatic annotation provides better accuracy results than purely automatic annotation since a user validates the proposed annotations. It is faster than manual annotation because the user chooses keywords from a proposed list. However,

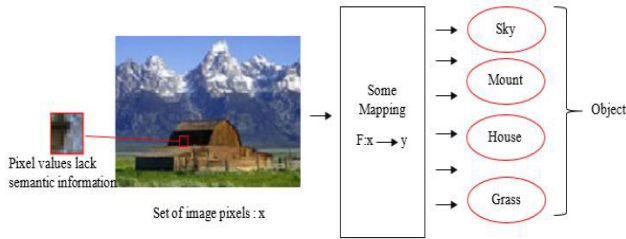


FIGURE 2. Example of the IA problem to find a mapping that can bridge the semantic gap between raw image pixels and semantic concepts like objects and scene categories

since it combines two types of annotation, it inherits the same drawbacks as manual and automatic annotation.

III. PROBLEM DEFINITION

The number of images generated daily by various internet websites and personal archives is continuously increasing, wherein the databases are attaining unimaginable sizes. The popularity of these large digital images' collections depends on their ease of utilization by internet users. However, not all of these image databases are often equipped with adequate indexing information so that any user can easily retrieve such images from anywhere on globe at any time. Presently, it isn't easy to access the image information through the Command interested by the user. Thus, an appropriate automated mechanism must be developed to efficiently characterize the image contents more meaningfully, instantly searching the information about the images required by the user. In the computer vision and multimedia domains, IA remains one of the most challenging problems, wherein the main aim is to create a map from a digital image to several labels. Creating such a map requires a clear understanding of the multifaceted semantic meaning based on the visual image contents. For example, the mountain image in Figure 2 containing some Sky, Mount, and a house post may be the appropriate objects to annotate, while the scene may be suitably labelled as "farm" or "view". It is feasible to utilize image annotation (IA) for the purpose of image retrieval, where images can be categorized and indexed according to their visual characteristics for future search purposes. In practice, a robust mechanism for such IA is lacking.

In computer science and engineering, the raw image pixels' inability to provide adequate devoid of ambiguity information, represent main IA's challenge which creating the semantic-level concepts. In addition, the text annotation suffers from various issues as the dictionary is well-defined and the syntax for combining alphabets into words and words into sentences is well-established, unlike text annotation, for which the dictionary relates words directly to the semantics [16] There no standard definition of 'words' or 'sentences' at present can be associated with the meaning of an image in the annotation. The aforementioned example (Figure 1) clearly illustrated the lack of a correlation between the pixels

and semantics that can be signified as the 'semantic gap.' Thus, it is necessary to resolve this semantic gap wherein the synergy of DL, DNNs, ML, ANNs and computer vision may be possible to embankment the existing gaps, thus developing called automated imaging association (IA) is a robust image classification scheme.

IV. IMAGE ANNOTATION APPROACH ARCHITECTURE

A. PROPOSED CNN ARCHITECTURE

Within the realm of artificial neural networks (ANNs), convolutional neural networks (CNNs) are harnessed for the purpose of capturing localized data characteristics. By assigning distinct weights to individual features, CNNs effectively streamline the overall complexity of the network. Due to their distinctive attributes, CNNs have garnered significant acclaim in the domain of pattern recognition [20]. For instance, a document-reading system is trained by jointly employing a CNN and a probabilistic model that incorporates language constraints. The architectural composition of a CNN encompasses three pivotal elements: 1) the input layer, 2) the hidden layer, and 3) the latent layer. Our proposed CNN architecture tuning the parameters of a neural network is a crucial aspect of optimizing its performance. The first convolutional layer applies 64 filters of size 3×3 to the input image. It uses the ReLU activation function, which introduces non-linearity to the network. The resulting feature maps have dimensions of $254 \times 254 \times 64$. The second convolutional layer applies 32 filters of size 3×3 to the output of the first layer. Again, ReLU activation is used. The feature maps produced by this layer have dimensions of $252 \times 252 \times 32$. The max pooling layer reduces the spatial dimensions by a factor of 2. It applies a 2×2 pooling window to the input and retains the maximum value within each window, resulting in feature maps of size $126 \times 126 \times 32$. The flatten layer reshapes the 3D output from the previous layer into a 1D vector. This step is necessary to connect the convolutional layers to the fully connected layers. Dropout is a regularization technique that helps prevent overfitting by randomly setting a fraction of input units to zero during training. The dropout rate is a hyperparameter that needs to be tuned. This dropout layer is applied to the output of the first dropout layer, which has been flattened. It further helps regularize the network. The first fully connected layer has 128 neurons and is connected to the output of the previous layer. The choice of activation function (ReLU) is a hyperparameter that can be tuned. The second fully connected layer serves as the output layer, these parameters define the architecture of the CNN.

B. FEATURES EXTRACTION FRAMEWORK

Figure 3 displays the flowchart of a typical features extraction framework. The image annotation (IA) operation comprised many laborers like the feature extraction procedure, suitable features in IA, selected mathematical transform for to set the feedback exercise, efficient features usage, effective features,

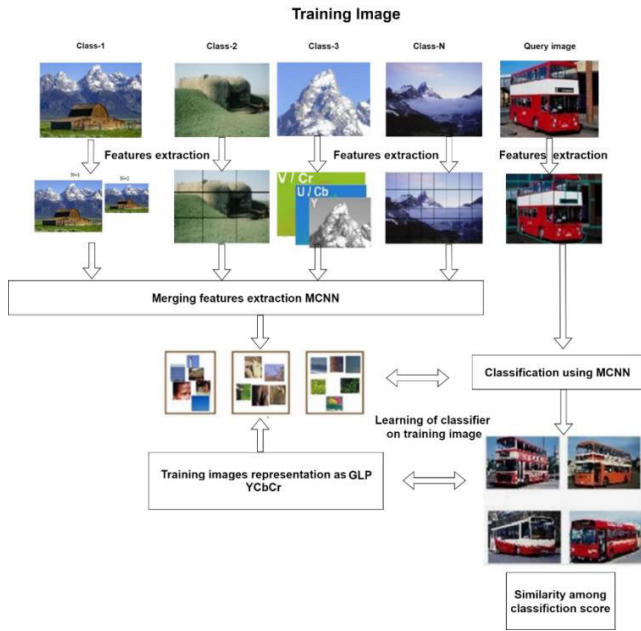


FIGURE 3. Flowchart of a typical features extraction framework

etc. An annotation system is considered effective if it can improve some characteristic factors.

In this view, low and high-level features of the image such as textures, shapes and colors were used for assembling all relevant information from the image to its recuperation.

C. YCBCR COLORCHANNEL

Various color models have demonstrated different levels of sensitivity in detecting signs of tampering in images. By and large, RGB and greyscale-based color systems are frequently utilized to identify forged images. However, a recent study by [17] found that employing chromatic channels instead of RGB or luminance can significantly enhance the effectiveness of detection. Figure 4 displays the proposed CNN-YCBCR color space model in this study. Figure 5 shows Color (RGB) images have three channels, Y, Cb and Cr. The YCbCr color model represents the colors in the form of the luminance (Y) and chrominance (Cb and Cr) components.

To calculate the Y, Cb, and Cr channels, the R, G, and B channels are used, Channels Eq. (1) was used. The proposed CNN-YCBCR model developed by combining the YCBCR color model and CNN model (Figure 5) could achieve improved image annotation.

$$\begin{pmatrix} Y \\ Cb \\ Cr \end{pmatrix} = \begin{pmatrix} 0.299 & 0.587 & 0.177 \\ -0.299 & -0.587 & 0.886 \\ 0.701 & -0.587 & -0.114 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} + \begin{pmatrix} 16 \\ 128 \\ 128 \end{pmatrix} \tag{4}$$

D. CNN-GAUSSIAN-LAPLACIAN PYRAMID

With the successive decomposition the image resolution becomes halved [18] due to the data generation sampling rate of 2.

Consequently, a pyramid of decomposed images yields the layers of decreasing resolution at the bottom layer wherein the original image occurs at the top (level 0). In order to reduce the horizontal and vertical resolution, First the input image is convoluted with a Gaussian low-pass filter before being down sampled by half. Essentially, the down-sampled filtered image becomes image (F1) at level 1 of the Gaussian pyramid. For each level in the Gaussian pyramid, the repeating process is described via the relation:

$$f = \sum_{m=-2}^2 \sum_{n=-2}^2 G(x,y) f_{i-1}(2i+m, 2j+n) \tag{5}$$

$$F_N = F_1 - D_1$$

$$F_N = F_N + F_1 \tag{6}$$

where G(x, y) is a low-pass filter, l is the level of the pyramid, N = 3 is the number of layers in the pyramid, 1 Cl 0 ≤ l ≤ N, i ≤ R, j ≤; Rl and Cl represent the corresponding number of rows and columns in Fl. The iterative image generation and boundary closure can be affected by the noise and dispersed sampling because each image in F1 has a different resolution. To complete the image enhancement process, one needs to retain the highest resolution image [19]. Furthermore, in order to keep the structural information from shifting to the next levels of the pyramid, the improved Retinex algorithm was first used. It enabled to recover the detail in images that was lost during the processing. The enhanced pyramid images were designated as ‘F1’. After constructing the image-enhanced Gaussian pyramid, a Laplacian pyramid was generated. The steps involved in the reconstruction were (i) an interpolated image of the topmost Gaussian pyramid (F1+1) results in the image D1 had the same resolution as the preceding image in the pyramid (‘F1’); (ii) a differential, a difference, FN, was stored in the Laplace residual set after D1 was subtracted from ‘F1 in order to yield ‘FN, which was used to reconstruct the image in the preceding layer; (iii) the reconstructed image was interpolated and the Laplacian was applied repeatedly until it Produced the same resolution as the original input. Based on the terminology described above, the following procedure can be described Figure 6 shows a multi-scale Gaussian–Laplacian pyramid derived from intermediate results of different pyramid levels. The sampling value for the pyramid in our test was 3 [20]. Figure 6 schematically presents the Gaussian-Laplacian image decomposition scheme. Figure 7 shows the intermediate results from different pyramid levels constructed using a Gaussian–Laplacian multi-scale pyramid. Column 1 shows the Gaussian pyramid image of the various pyramid levels and the images of pyramids at varying levels of enlargement are displayed in Column 2. Processed images of the Laplace reconstruction are displayed in Column 3 and the final image is in the first picture of Column 3.

E. IMAGE ANNOTATION (IA) SCHEME

The final image obtained from the Laplacian multi-scale pyramid was entered into CNN with multi-convolution layer

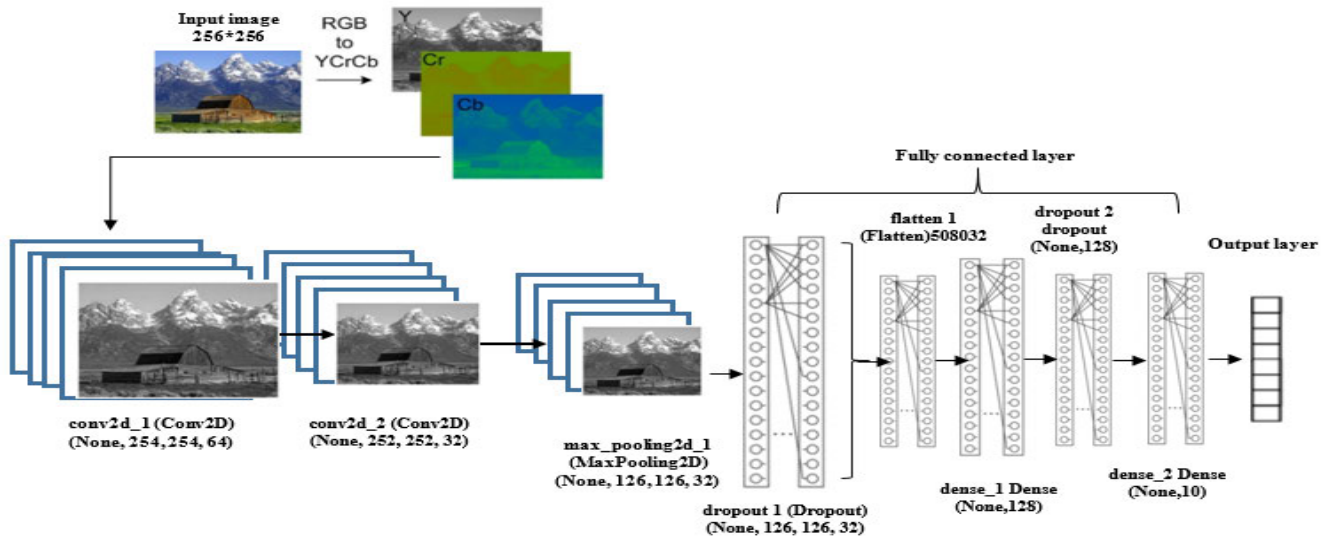


FIGURE 4. CNN-YCBCR color space model.

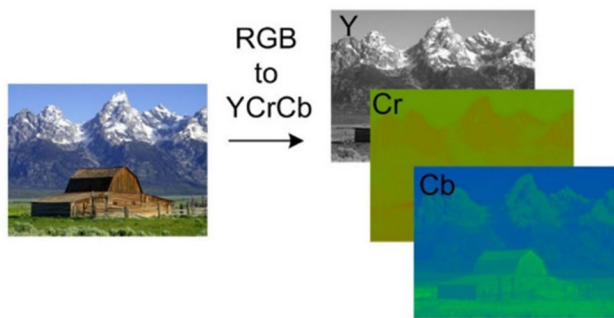


FIGURE 5. Convert RGB to YCbCr images.

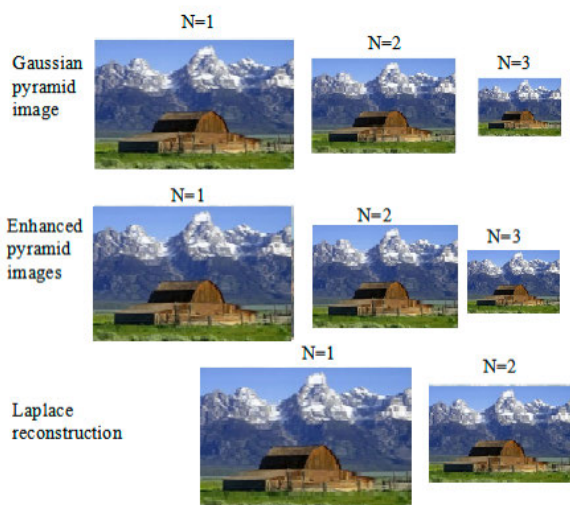


FIGURE 6. Schematic presentation of the Gaussian-Laplacian image.

and multi-max pooling layer to extract the features more clearly. This allowed achieving a high performance of the

retrieved image as described in the deep learning and architectures of the CNN model [21]. In the realm of artificial intelligence, the process involves automatically matching terms from a dictionary to an image. The input is the image that needs to be described, and the output is a collection of relevant terms that effectively depict the image. While a computer can analyze basic attributes such as colors, textures, and shapes, it still requires assistance to comprehend these features in a meaningful manner. In contrast, humans can swiftly interpret the meaning of an image. Consequently, the primary challenge in artificial intelligence lies in overcoming the “semantic gap” between the computable low-level features and human understanding of images. This particular issue has recently garnered significant attention in AI research. Addressing these difficulties, researchers have come up with various models to tackle this problem and close the semantic gap more effectively.

1) PROPOSED IA ARCHITECTURE

A comprehensive overview of the proposed IA architecture is provided in this section. Three main stages made up the implementation of the planned IA system. First, the system’s training phase, which made use of a database of annotated images, was crucial. Second, the annotated image was produced using the trained system’s work on the raw data. Finally, image retrieval was done to gauge how well the suggested IA system performed. In the first training phase, CNN automatically extracted features from the standard training database. By comprehending the contents of the photos, the automated features extraction procedure easily created the feature vector. The second step produced the model for annotation, which then annotated the fresh images by modelling the features using a learning process. The input for the second phase was the image without any annotations. In the previous

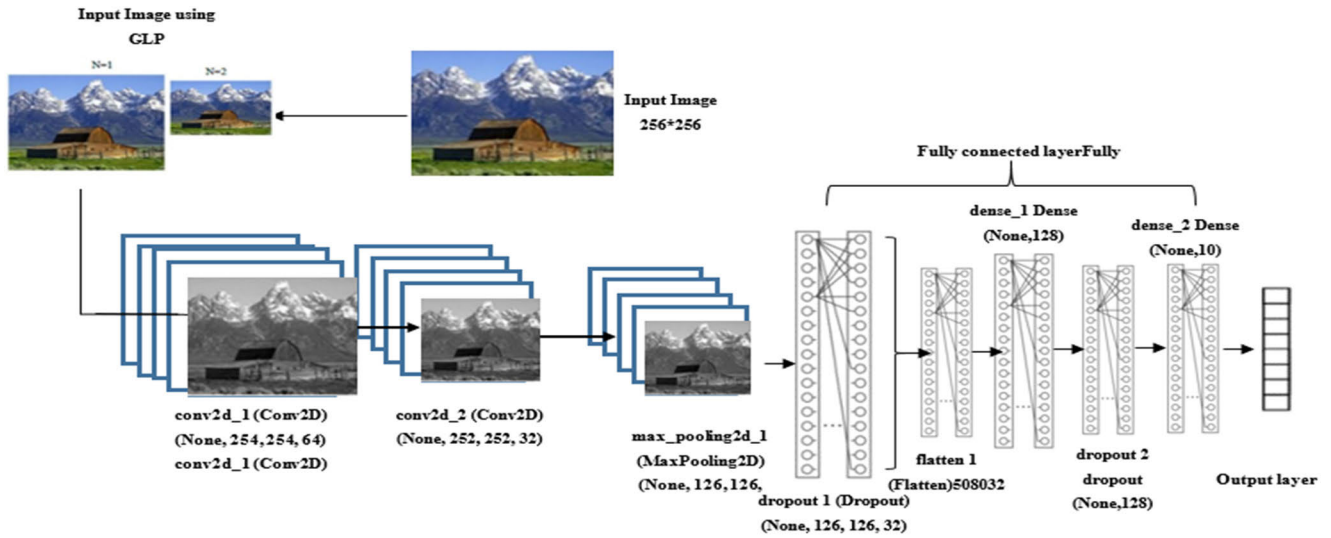


FIGURE 7. The proposed CNN-Gaussian-Laplacian pyramid model.

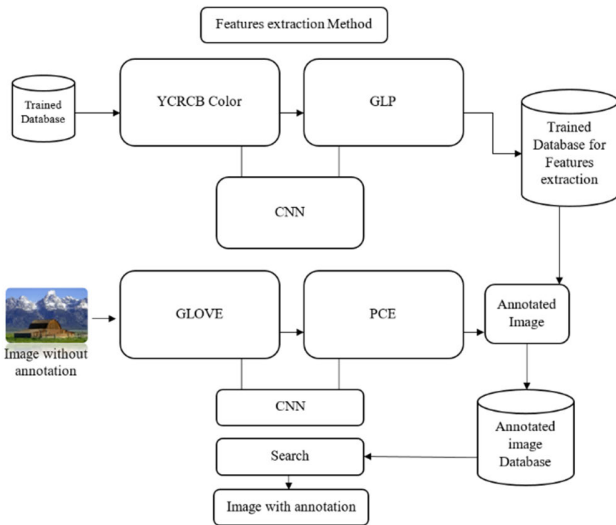


FIGURE 8. The proposed architecture of IA scheme.

Phase, after developing the annotation model, the next step was to extract features to generate the visual attributes for the contents. The image was given a correct semantic label according to the contents of the model created in the previous step. As a consequence, the output was an image with annotations. The third phase used the annotation phase’s image files as the data storage. After issuing a textual query, the system generated a list of relevant photos. The annotation made retrieving photographs based on their content simpler and more accurate. The suggested IA system design is shown in Figure 8.

2) GLOBAL VECTORS (GLOVE) FOR WORD REPRESENTATION

The GloVe algorithm is used to find the vector representations of the words via the unsupervised learning. Based

on the global co-occurrence statistics from a corpus, the representations of word vector space are trained based on the word co-occurrence statistics [22]. The resulting representations encapsulate many interesting linear substructures of the words as described below.

F. NEAREST NEIGHBORS

A measure of linguistic or semantic similarity between two words vectors can be defined. A cosine similarity between the vectors can be calculated as the Euclidean distance (or distance between Euclides). These metrics can reveal words that are rare but relevant and would be entirely outside the scope of the average human [23].

G. LINEAR SUBSTRUCTURES

Two words can be compared in terms of their correlation using the nearest neighbor metric. The more complex the relationship, the more difficult it is to convey with a single number, so it is sometimes advantageous to simplify. For instance, both the words man and woman describe human beings. However, these terms are often considered opposite to each other since they highlight a primary axis along which humans differ from one another. Thus, in order to capture a quantitative sense where the nuance required to differentiate between man and woman can be captured by a model having more than one number for each pair of words. The difference between two-word vectors that makes an ideal candidate for an enlarged set of discriminative numbers [18]. The purpose of GloVe is to capture as much of the meaning supplied by the juxtaposition of two words as possible by using vector differences. In addition to sex or gender, there are various other concepts that distinguish man from woman such as brother and sister, king and queen, and so forth. In mathematics, one might expect that the vector differences between a man and a woman, a king and a queen, and a

brother and a sister would be roughly equal. The above set of visualizations shows this property as well as other interesting patterns. The GloVe model is a log-bilinear model with a weighted least squares objective. The model based on the observation of ratios for word-to-word co-occurrence probabilities of encoded meaning. For example, co-occurrence probabilities of ice and steam with various probe words in the corpus. Based on 6 billion words in a corpus some probability estimates are given.

H. COMBINING THE GLOVE WITH PRINCIPAL COMPONENT ANALYSIS (PCA)

Despite having a high-dimensional representation of words, a projection onto a lower dimension is required to better grasp the features' vectors and increase performance. This effort aimed to improve comprehension and disclose data structures, which are crucial for language modeling and word prediction. Assume N is the corpus's word count and d is the length of each word features vector. The features' covariance matrix R ($d \times d$) was used to decrease it to k (where $k < d$) [24]. Thus, the covariance between two features j and l was calculated as:

$$\sigma_{jl} = \frac{1}{n-1} \sum_{i=1}^N (x_{ij} - \mu_j)(x_{il} - \mu_l) \quad (7)$$

where x_{ij} is the j th feature of the word i . The matrix R was generated using:

$$R = \frac{1}{n-1} (X - \vec{\mu})^T (X - \vec{\mu}) \quad (8)$$

The covariance matrix R calculated eigenvalues and eigenvectors to project the original data feature matrix of size ($N \times d$) using the best k eigenvectors as axes. The k largest eigenvalues of a projection matrix P ($d \times k$) were selected as eigenvectors. The projection matrix was multiplied by the data features' matrix to create an $N \times k$ matrix. This reduction may preserve key features and data [11]. G was the glove and

W was Word2vec. Word to vector space with insertion embedding. Both embeddings employed different vector spaces. Multiplying a rotation matrix by G aligns it with W . Random matrix and gradient descent decreased reconstruction errors.

V. IMPLEMENTATION OF IMAGE ANNOTATION

The proposed IA model simulated on a public DL software called Keras [28] based on Tensorflow [29]. Based on Keras settings, the weights of the NNs were set. The deep network layers were all set jointly with ADADELTA [30]. The complete network was trained on a Dell T1900 CPU equipped with 32GB of memory. This novel DL system was assessed for its computing classification accuracy using the procedure described in previous section. The suggested convolutional neural network (CNN) model, illustrated in Figure 9, represents the envisioned CNN setup. It incorporates the Y, Cb, and Cr color channels, along with GLP as the input features derived from images. This configuration is implemented using the Keras library. It used two convolutional layers, one

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, 256, 256, 1)	0	
input_2 (InputLayer)	(None, 256, 256, 1)	0	
input_3 (InputLayer)	(None, 256, 256, 1)	0	
conv2d_1 (Conv2D)	(None, 254, 254, 64)	640	input_1[0][0]
conv2d_3 (Conv2D)	(None, 254, 254, 64)	640	input_2[0][0]
conv2d_5 (Conv2D)	(None, 254, 254, 64)	640	input_3[0][0]
conv2d_6 (Conv2D)	(None, 252, 252, 32)	18464	conv2d_5[0][0]
max_pooling2d_1 (MaxPooling2D)	(None, 126, 126, 32)	0	conv2d_2[0][0]
max_pooling2d_2 (MaxPooling2D)	(None, 126, 126, 32)	0	conv2d_4[0][0]
dropout_1 (Dropout)	(None, 126, 126, 32)	0	max_pooling2d_1[0][0]
dropout_2 (Dropout)	(None, 126, 126, 32)	0	max_pooling2d_2[0][0]
flatten_1 (Flatten)	(None, 508032)	0	dropout_1[0][0]
flatten_2 (Flatten)	(None, 508032)	0	dropout_2[0][0]
flatten_3 (Flatten)	(None, 508032)	0	dropout_3[0][0]
concatenate_1 (Concatenate)	(None, 1524096)	0	flatten_1[0][0] flatten_2[0][0] flatten_3[0][0]
dense_1 (Dense)	(None, 1200)	1828916400	concatenate_1[0][0]
dense_2 (Dense)	(None, 400)	480400	dense_1[0][0]
dropout_4 (Dropout)	(None, 400)	0	dense_2[0][0]
dense_3 (Dense)	(None, 10)	4010	dropout_4[0][0]
Total params: 1,829,458,122			
Trainable params: 1,829,458,122			
Non-trainable params: 0			

FIGURE 9. The proposed CNN-GLP configuration.

max pooling layer, one full connected layer with two drop-out and flatten layers, and two dense layers to train all 65,048,618 model parameters.

VI. EXPERIMENTAL PROCEDURES

The proposed IA system was implemented using Python with the DL framework Keras [28] and TensorFlow [29]. As a wrapper to write all activities through the dataset testing and training the Python coding was used. First, Python 3.5 was installed in the set environment. Next, TensorFlow and PrettyTensor were installed as package in Python. PrettyTensor provided much simpler ways to construct the NNs in TensorFlow, thus enabling to focus on the design model instead of low-level implementation details. The CNN-GLP model's quality and reliability were evaluated by employing three standard datasets as benchmarks. The comprehensive annotation framework of CNN was utilized to address the newly developed IA problems. To maximize the effectiveness of the novel IA framework, the system integrated and fused the extracted features with the CNN's architecture. First, it explained the datasets and evaluated the metrics. Second, for each methods the results were analyzed and presented shortly for further validation. Finally, a comparison was made between the proposed model performance and different other state-of-the-art methods of image annotation. In addition, several examples were provided to explain the detailed working principle of the image annotation process.

VII. DATABASE

The experiments were conducted on three most popular IA databases (Table 1) Corel-5K [26], ESP-Game [19] and

TABLE 1. Detailed information of each database.

Database	Number of Images	Vocabulary Size	Training Size	Testing Size	Words per Image	Images per Word
Corel-5K (2007)	5 000	260	4 500	500	3.4	58.6
ESP-Game (2003)	20 770	268		2 081	4.7	362.7
IAPRTC-12(2010)	19 627	291	17 665	1 962	5.7	347.7

IAPRTC-12 [27]. The Corel-5K dataset was created and released in 2007 [26], the ESP Game itself was launched in 2003, and IAPRTC-12 dataset was released in 2010.

VIII. PERFORMANCE EVALUATION CRITERIA

The effectiveness of the enhanced of (IA) system was evaluated by implementing it with a standard dataset. This paper examined several cutting-edge works in order to measure its performance. Various metrics, including recall, accuracy, F measure, and N+, were used to assess the performance of the IA. The annotation rate was also considered as a measure of evaluation [28].

IX. EXPERIMENTS RESULTS

A. EXPERIMENTS RESULTS OF FEATURES EXTRACTION

The performance metrics of features extraction techniques such as YCBCR and GLP transform were assessed. Table 2 shows the classification results achieved by these descriptors when applied three datasets. Certainly, the CNN-GLP transform performed better than other descriptors in terms of the precision and accuracy. For the training and testing IAPRTC-12 dataset the achieved classification accuracy was approximately 95% and 92%, respectively. The CNN-GLP scheme also achieved 91% of classification accuracy for the Corel-5K dataset. The YCBCR-CNN scheme accomplished about 87% of classification accuracy When tested on IAPRTC-12 dataset, whereas CNN-GLP scheme attained as much as 92% of classification accuracy when implemented on Corel-5K dataset. Overall, the experimental evaluations of all three schemes revealed very encouraging outcomes when applied on the Corel5k, ESP Game, and IAPRTC-12 datasets. The accuracy performance of the proposed approach was tested on IAPRTC-12 dataset and the following observations were made. The accuracy of CNN-GLP approach was highest (92%), followed by the YCBCR-CNN (87%) scheme. In terms of performances, YCBCR-CNN schemes achieved the best classification when compared with other conventional methods (Table 2 and 3). The accuracy performance of the developed schemes when implemented on ESP Game was quite encouraging. The accuracy of CNN-GLP approach was highest (91%) followed by the YCBCR-CNN (71%) approach. The image features classification performances of the CNN-GLP scheme were the best compared to several other reported techniques (Table 2). The accuracy performance for IAPRTC-12 dataset can be summarized as

TABLE 2. The of the retrieved images obtained classification accuracy using ycbcr-cnn and glp-cnn schemes when compared with the existing methods.

Database	Number of Samples	Features Type	Accuracy (%)	
			Training	Testing
Corel-5K	Train 4000 Test 1000	YCBCR-CNN	91	87
		GLP-CNN	95	89
ESP-Game	Train 17 689 Test 3 081	YCBCR-CNN	94	71
		GLP-CNN	97	91
IAPRTC-12	Train 16 665 Test 2 962	YCBCR-CNN	98	81
		GLP-CNN	95	92

follows: the accuracy of CNN-GLP approach was the highest (92%) followed by the YCBCR-CNN (81%) scheme. The classification performances of the CNN- GLP was the best compared to the existing schemes (Table 2).

B. EXPERIMENTS RESULTS OF IA

Assessing the effectiveness of the proposed IA approach through evaluations on popular image annotation databases: Corel-5K, ESP-Game, and IAPRTC-12. A total of 1000 images from Corel-5K dataset, 3081 images from ESP-Game and 2962 from IAPRTC-12 were tested. One of the main motivations for the IA is to achieve annotation-based image retrieval. Therefore, it appeared natural to use the retrieval result to reflect the performance of the developed IA system. Majority of the IA-related research measures the accuracy and recall through the process of retrieving testing images with individual keyword. In this view, present work computed N+, the number of tags that are correctly assigned to at least one test image. Table 3 shows the experimental findings of IA-CNN-GLP model implementation on three popular databases. The trained translation table was used to annotate the images for the system performance evaluation. Table 3 displays the higher evaluation results. The values of precision, recall, F, and N+ for Corel-5K, ESP-Game and IAPRTC-12 datasets were correspondingly 40, 48, 43, and 200; 39, 35, 41, and 210 and 49, 45, 42, and 260. The IA was more frequent for both training and testing when applied on Corel dataset and performed better compared to ESP-Game and IAPRTC-12 dataset [29]. As one might expect ice to form more frequently when solids are present than when gases are present, similarly, steam forms more frequently when gases are present than when solids are present.

Water appears frequently with both words, and fashion appears infrequently with both words. The relationship between the probabilities and noise cancels out because no discriminative words like water and fashion have large values

TABLE 3. Experimental findings of ia implementation on three popular datasets.

Database	Method	Evaluation			
		P	R	F	N+
Corel-5K	IA-CNN-GLP	40	48	43	200
ESP-Game	IA-CNN-GLP	39	35	41	210
IAPRTC-12	IA-CNN-GLP	49	45	42	250

(much greater than 1) that correlate well with the properties specific to ice, and smaller values (much less than 1) that correlate well with the properties specific to steam. By encoding some crude meaning associated with the abstract concept of thermodynamic phase, the probability ratio can encode some form of meaning associated with it [37].

X. COMPARING THE PERFORMANCE OF THE PROPOSED METHODS WITH ALTERNATIVE APPROACHES

Table 4 displays a comparison of the experimental outcomes attained by utilizing the suggested image annotation system on three distinct datasets. The image annotation system recommended was executed through widely available deep learning software, namely Keras [28] and Tensorflow [29]. Keras initialized neural network weights. ADADELTA started all deep network layers concurrently [30]. The 32-GB Dell Precision T1900 CPU system educated the whole Network. Section IV examined DL system calculating classification accuracy. Table 4 shows the suggested CNN setup utilizing Kears library’s Y, Cb, and Cr image color channels. [38] The suggested CNN model’s average accuracy, recall, and F-measure for each dataset were compared to literature (Figure 11, 12, 13). The IA-MCNN-GLP technique showcased superior performance when compared to 2PKNN, SEM, and GAN. It proved to be more suitable for annotation tasks, as it exhibited higher precision and recall rates. espGame and laprtc12 improved recall and F-measured for the planned IA system. The novel IA system, which relies on autonomous feature extraction and object learning representation, demonstrated superior performance on all three datasets by delivering the most effective features. As shown in Figure 10, the CNN-GLP model is accurate on both training and testing datasets. It was not possible to exceed the accuracy of the training set over the accuracy of the testing set. The suggested IA system has the greatest F-value, suggesting its efficacy and resilience. Table 4 shows manual labeling and prediction examples from training and testing datasets. The suggested IA approach expanded labels in the training dataset using the original pictures [5] with less labels. Labels may also stay on images. By applying the recommended approach to evaluate the testing subset, the effectiveness of

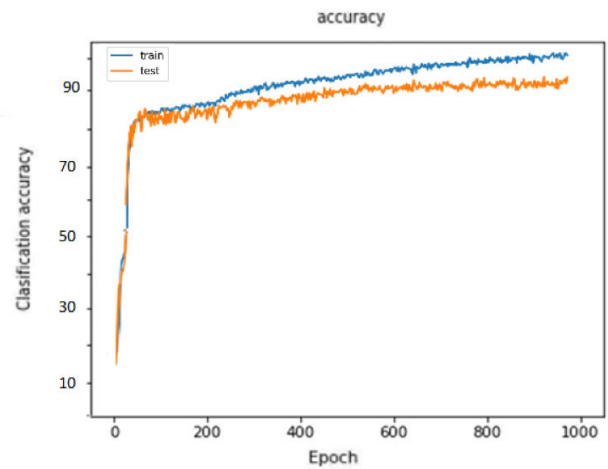


FIGURE 10. Classification accuracy curve for CNN-GLP in Corel-5k database.

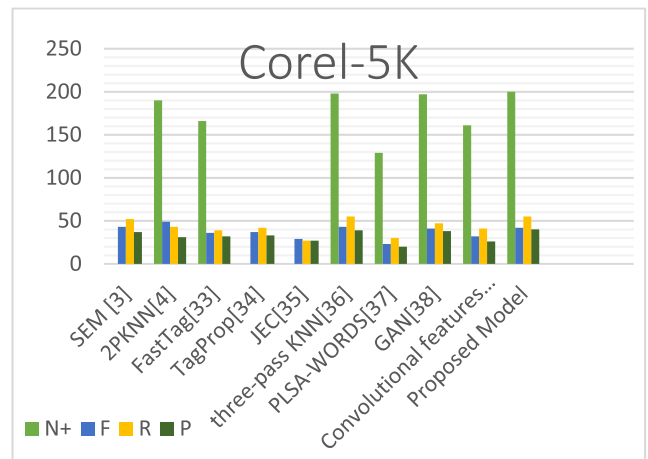


FIGURE 11. The outcomes achieved through the suggested method applied to three datasets (Corel-5k) will now be compared.

each dataset is measured and assessed in Corel-5k follows the trend of: CNN-GLP approach achieved highest P (40), GAN was 38; KNN approach produced the highest R (55); CNN-GLP approach did not achieve highest F1 (43); the highest F1 was 49 and the difference between them was 6. CNN-GLP’s N+ (200) was greater than the other five algorithms’ and improved by at least 3. CNN-GLP and three-pass KNN were the best IA approaches for annotation. The suggested IA technique performed best on the ESP-Game dataset: 2PKNN (40) and CNN-GLP (38) had the greatest P. CNN-GLP has the greatest R (46) and F1 (50) compared to SEM (42) and all other algorithms (49). CNN-GLP generated the greatest N+ with three-pass KNN (260) compared to other algorithms (259), improving by at least 1. Table 4 showed CNN-GLP and SEM had the highest N+.

The CNN-GLP technique surpassed the 2PKNN in annotation performance, since the difference in F1 was 1. [39], [40] It was claimed that the proposed IA system attained optimum annotation performance with the highest P value of

TABLE 4. Presents the experimental results achieved by the proposed IA system on three datasets, in comparison with other existing works.

Dataset	Corel-5k				ESP-Game				IAPRTC-12			
	P	R	F	N+	P	R	F	N+	P	R	F	N+
Measure												
SEM [3]	37	52	43		38	42	40	258	41	39	40	284
2PKNN [4]	31	43	49	190	40	39	49	255	37	41	46	279
FastTag [33]	32	39	36	166	39	35	29	247	32	34	33	280
TagProp [34]	33	42	37		37	27	32		46	35	40	
JEC [35]	27	27	29		22	25	23		28	29	28	
KNN [36]	39	55	43	198	37	35	40.8	259	51	37	37	278
PLSA [37]	20	30	23	129	20	24	21	201	23	25	23	207
GAN [38]	38	47	41	197	-	-	-	-	44	38	43	199
Convolutional [39]	26	41	32	161	37	33	37	258	46	32	38	258
Proposed Model	40	48	43	200	38	46	50	260	42	41	39	280

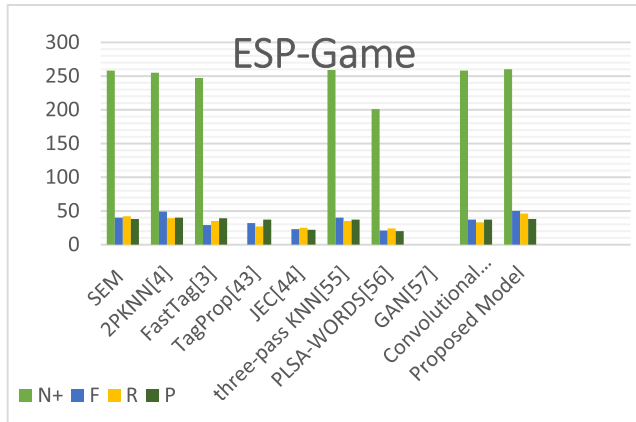


FIGURE 12. The outcomes achieved through the suggested method applied to three datasets (ESP-Game) will now be compared.

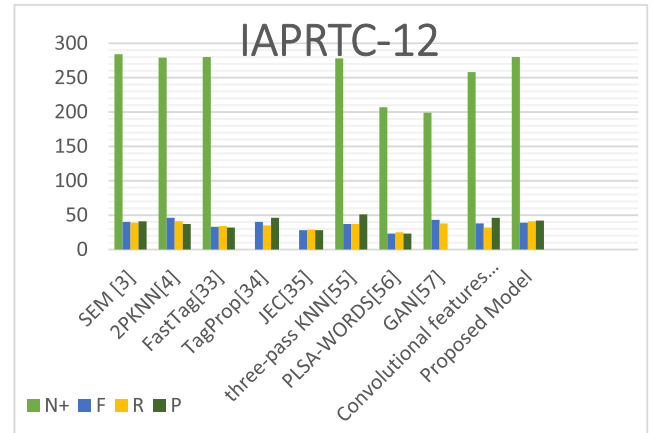











FIGURE 13. The outcomes achieved through the suggested method applied to three datasets (IAPRTC-12) will now be compared.

46 for IAPRTC-12, 42 for CNN-GLP and lowest for PLSA [41] Table 5 highlights the disparities in keyword generation between our novel model and previous techniques. Our newly proposed IA system demonstrates enhanced resilience, accurately labeling the majority of photographs without necessitating an excessive number of labels. This underscores the strength of our system. Unlike earlier systems, [41] the current IA system solves the issue of a fixed number of labels and nearby labels in the image, making label identification simpler. A limited number of photos may prevent the model from properly learning the features, resulting in an image class with inaccurate labels or fewer labels. Due to the fixed number of labels, standard SEM, E2E-DCNN, and CNN-THOP have several over-labeling and under-labeling issues, making exact labeling difficult. In summary, our research showed that the suggested IA approach was more

accurate and effective than all prior state-of-the-art methods [42]. The proposed IA (Image Annotation) system represents promising results and advantages over existing methods. However, its limitations should be considered to ensure a comprehensive understanding of the system’s applicability and potential enhancements. Even though we emphasize our approach’s computational efficiency, it still requires a significant amount of computing power, especially when handling large datasets. There may be limitations to this for researchers that lack high-performance computing facilities. It might still be challenging to interpret these features, despite our approach’s combination of high-level and low-level features. It is inevitable that there will be annotation errors with our automated annotation system. There is a possibility that images may be mislabeled or annotated inaccurately in certain cases, which can lead to consequences in critical applications.

TABLE 5. Provides examples of dataset utilized by the proposed IA system.

Datasets	Proposed Method	SEM [3] Annotation	CNN [39] Annotation
Corel-5k			
Manual Labelling	Beach, People, Rocks, Kauai	Building, Bus, Tree	Elephant, River, Tree
Prediction	Beach, Person, Rocks, Kauai, Water, Clouds	Bus, Building, Tree, Sky, House, Land	Elephant, Tree, Land, River, Sky
ESP-GAME			
Manual Labelling	Forest, Mountain, River, Sky, Tree, Water	Stone, White, Cow, Dirt, Tail, Bull, Grass,	Person, Sky, River, Lawn, Road
Prediction	Forest, Mountain, River, Sky, Tree, Water, Land	Stone, White, Cow, Dirt, Tail, Bull, Grass, Land, Wall	Grass, Sky, River, Lawn, Road, Girl, Slut
IAPRTC-12			
Manual Labelling	City, House, Roof, Sky, Valley, View	Fountain, Tree	Grandstand, Lawn, Player, Roof, Round, Stadium
Prediction	City, House, Roof, Sky, Valley, View, Tree, Clouds	Tree, Sky, Building, Cloud, Lawn, Land	Grandstand, Lawn, Player, Roof, Round, Stadium, Lights, People

XI. CONCLUSION

This study presented CNN features and neighbors to describe images using the GLP Transform and YCBCR color. The term “sea” is used to describe the process of a person’s re-entry into a domain. To increase proposed IA system reliability and efficiency, all image information was built utilizing high-level and low-level features such forms, textures, and colors. Features extraction of automatic image as linked to distributed impersonation methods like encoding and storing. CNN designs have substantial impacts on diverse datasets. The CNN-GLP approach not only outperforms established algorithms across multiple datasets, demonstrating its innovation and effectiveness through superior precision (P), recall (R), and F1-score (F1) performance, but it also serves as a bridge across the semantic gap. By achieving higher precision and recall rates compared to existing methods like 2PKNN, SEM, and GAN, this approach firmly establishes its suitability for annotation tasks and its capability to bridge the semantic gap in image description. Three datasets showed that the image annotation model worked. The experimental results for three public datasets—Corel-5K, ESP-Game, and Iaprtc-12—showed that the CNN-GLP had 40, 38, and 42 precisions, 55, 46, and 41 recall, and 42, 50, and 39 F1 values. N+ was 200, 260, and 280. The ultimate IA system balances precision, recall, and accuracy. GLP balanced accuracy and recall and chose the best features for the suggested IA approach. In upcoming research, improvements to the CNN-GLP model can be pursued through exploration of various architectures, hyperparameter adjustments, and training methods. Furthermore, advanced deep learning methods can be explored to deepen the system’s understanding of semantics. Using transfer learning to improve performance when there is limited labeled data, fine-tune pre-trained models for improved performance. When using huge training datasets, [47] several image annotation methods were computationally costly due to their sophisticated training process. Thus, this approach is computationally efficient, reliable, and accurate.

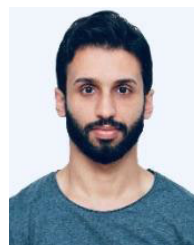
ACKNOWLEDGMENT

The authors would like to thank the AIDA Laboratory, College of Computer and Information Sciences (CCIS), Prince Sultan University, Riyadh Saudi Arabia, for its support in publishing this research.

REFERENCES

- [1] F. Baig, Z. Mehmood, M. Rashid, M. A. Javid, A. Rehman, T. Saba, and A. Adnan, “Boosting the performance of the BoVW model using SURF-CoHOG-based sparse features with relevance feedback for CBIR,” *Iranian J. Sci. Technol., Trans. Electr. Eng.*, vol. 44, no. 1, pp. 99–118, Mar. 2020.
- [2] M. Tzelepi and A. Tefas, “Deep convolutional learning for content based image retrieval,” *Neurocomputing*, vol. 275, pp. 2467–2478, Jan. 2018, doi: 10.1016/j.neucom.2017.11.022.
- [3] M. S. Haji, M. H. Alkawaz, T. Saba, and A. Rehman, “Content-based image retrieval: A deep look at features prospectus,” *Int. J. Comput. Vis. Robot.*, vol. 9, no. 1, pp. 14–38, 2019.
- [4] Q. Cheng, Q. Zhang, P. Fu, C. Tu, and S. Li, “A survey and analysis on automatic image annotation,” *Pattern Recognit.*, vol. 79, pp. 242–259, Jul. 2018.

- [5] J. Tang, X. Shu, Z. Li, Y.-G. Jiang, and Q. Tian, "Social anchor-unit graph regularized tensor completion for large-scale image retagging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 2027–2034, Aug. 2019.
- [6] T. Gui, R. Ma, Q. Zhang, L. Zhao, Y.-G. Jiang, and X. Huang, "CNN-based Chinese NER with lexicon rethinking," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 4982–4988.
- [7] N. Abbas, T. Saba, D. Mohamad, A. Rehman, A. S. Almazyad, and J. S. Al-Ghamdi, "Machine aided malaria parasitemia detection in Giemsa-stained thin blood smears," *Neural Comput. Appl.*, vol. 29, no. 3, pp. 803–818, Feb. 2018.
- [8] U. Sharif, Z. Mehmood, T. Mahmood, M. A. Javid, A. Rehman, and T. Saba, "Scene analysis and search using local features and support vector machine for effective content-based image retrieval," *Artif. Intell. Rev.*, vol. 52, no. 2, pp. 901–925, Aug. 2019.
- [9] T. Saba, S. T. F. Bokhari, M. Sharif, M. Yasmin, and M. Raza, "Fundus image classification methods for the detection of glaucoma: A review," *Microsc. Res. Technique*, vol. 81, no. 10, pp. 1105–1121, Oct. 2018.
- [10] M. Mundher, D. Muhamad, A. Rehman, T. Saba, and F. Kausar, "Digital watermarking for images security using discrete slantlet transform," *Appl. Math. Inf. Sci.*, vol. 8, no. 6, pp. 2823–2830, Nov. 2014.
- [11] M. Yousuf, Z. Mehmood, H. A. Habib, T. Mahmood, T. Saba, A. Rehman, and M. Rashid, "A novel technique based on visual words fusion analysis of sparse features for effective content-based image retrieval," *Math. Problems Eng.*, vol. 2018, pp. 1–13, Jan. 2018.
- [12] I. M. Hameed, S. H. Abdhussain, and B. M. Mahmood, "Content-based image retrieval: A review of recent trends," *Cogent Eng.*, vol. 8, no. 1, Jan. 2021, Art. no. 1927469.
- [13] K. Bharathi and M. C. Mohan, "A deep learning approach for content-based image retrieval using sparse auto-encoder," *Turkish J. Comput. Math. Educ.*, vol. 13, no. 3, pp. 27–32, 2022.
- [14] M. Alrahhah and K. P. Supreethi, "Multimedia image retrieval system by combining CNN with handcraft features in three different similarity measures," *Int. J. Comput. Vis. Image Process.*, vol. 10, no. 1, pp. 1–23, Jan. 2020.
- [15] T. Saba, A. Rehman, Z. Mehmood, H. Kolivand, and M. Sharif, "Image enhancement and segmentation techniques for detection of knee joint diseases: A survey," *Current Med. Imag. Rev.*, vol. 14, no. 5, pp. 704–715, Sep. 2018.
- [16] A. Sarwar, Z. Mehmood, T. Saba, K. A. Qazi, A. Adnan, and H. Jamal, "A novel method for content-based image retrieval to improve the effectiveness of the bag-of-words model using a support vector machine," *J. Inf. Sci.*, vol. 45, no. 1, pp. 117–135, Feb. 2019.
- [17] G. Zhang, C.-H.-R. Hsu, H. Lai, and X. Zheng, "Deep learning based feature representation for automated skin histopathological image annotation," *Multimedia Tools Appl.*, vol. 77, no. 8, pp. 9849–9869, Apr. 2018.
- [18] M. M. Adnan, M. S. M. Rahim, A. R. Khan, T. Saba, S. M. Fati, and S. A. Bahaj, "An improved automatic image annotation approach using convolutional neural network-slantlet transform," *IEEE Access*, vol. 10, pp. 7520–7532, 2022.
- [19] Y. Ma, Y. Liu, Q. Xie, and L. Li, "CNN-feature based automatic image annotation method," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3767–3780, Feb. 2019.
- [20] A. Rehman, "An ensemble of neural networks for nonlinear segmentation of overlapped cursive script," *Int. J. Comput. Vis. Robot.*, vol. 10, no. 4, pp. 275–288, 2020.
- [21] V. N. Murthy, A. Sharma, V. Chari, and R. Manmatha, "Image annotation using multi-scale hypergraph heat diffusion framework," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2016, pp. 299–303, doi: 10.1145/2911996.2912055.
- [22] Z. Lingxin, S. Junkai, and Z. Baijie, "A review of the research and application of deep learning-based computer vision in structural damage detection," *Earthq. Eng. Eng. Vib.*, vol. 21, no. 1, pp. 1–21, Jan. 2022.
- [23] Y. Verma and C. V. Jawahar, "Image annotation using metric learning in semantic neighbourhoods," in *Proc. Eur. Conf. Comput. Vis.* (Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7574, 2012, pp. 836–849.
- [24] F. M. Rammoo and M. N. Al-Hamdani, "Detecting the speaker language using CNN deep learning algorithm," *Iraqi J. Comput. Sci. Math.*, vol. 3, no. 1, pp. 43–52, Jan. 2022.
- [25] Z. Jiahao, Y. Jiang, R. Huang, and J. Shi, "EfficientNet-based model with test time augmentation for cancer detection," in *Proc. IEEE 2nd Int. Conf. Big Data, Artif. Intell. Internet Things Eng. (ICBAIE)*, Mar. 2021, pp. 548–551.
- [26] F. Gao, S. Ji, J. Guo, Q. Li, Y. Ji, Y. Liu, S. Feng, H. Wei, N. Wang, and B. Yang, "ID-Net: An improved mask R-CNN model for intrusion detection under power grid surveillance," *Neural Comput. Appl.*, vol. 33, no. 15, pp. 9241–9257, Aug. 2021.
- [27] T. Uricchio, L. Ballan, L. Seidenari, and A. Del Bimbo, "Automatic image annotation via label transfer in the semantic space," *Pattern Recognit.*, vol. 71, pp. 144–157, Nov. 2017.
- [28] F. Chollet, "Keras documentation," Keras.Io, Packt Publishing Ltd, Livery Place, Tech. Rep. 3, 2015. [Online]. Available: <https://keras.io>
- [29] R. Bibi, Z. Mehmood, R. M. Yousaf, T. Saba, M. Sardaraz, and A. Rehman, "Query-by-visual-search: Multimodal framework for content-based image retrieval," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 11, pp. 5629–5648, Nov. 2020.
- [30] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," 2012, *arXiv:1212.5701*.
- [31] J. Cao, A. Zhao, and Z. Zhang, "Automatic image annotation method based on a convolutional neural network with threshold optimization," *PLoS ONE*, vol. 15, no. 9, Sep. 2020, Art. no. e0238956.
- [32] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," *Int. J. Comput. Vis.*, vol. 90, pp. 88–105, Oct. 2010.
- [33] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamsad, C. Kern, J. R. Ledsam, M. K. Schmid, K. Balaskas, E. J. Topol, L. M. Bachmann, P. A. Keane, and A. K. Denniston, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis," *Lancet Digit. Health*, vol. 1, no. 6, pp. e271–e297, Oct. 2019.
- [34] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
- [35] R. Sinhal and I. A. Ansari, "Comparative analysis of watermark reconstruction using discrete wavelet transform and slantlet transform for user identification in social media," in *Proc. IEEE Int. Students Conf. Elect., Electron. Comput. Sci. (SCEECS)*, Feb. 2020, pp. 1–5.
- [36] T. Saba, A. Rehman, A. Altameem, and M. Uddin, "Annotated comparisons of proposed preprocessing techniques for script recognition," *Neural Comput. Appl.*, vol. 25, no. 6, pp. 1337–1347, Nov. 2014.
- [37] C. Harris, "ClueMeIn: Obtaining more specific image labels through a game," in *Proc. Workshop Games Natural Lang. Process.*, May 2020, pp. 10–16.
- [38] H. Younis, M. H. Bhatti, and M. Azeem, "Classification of skin cancer dermoscopy images using transfer learning," in *Proc. 15th Int. Conf. Emerg. Technol. (ICET)*, Dec. 2019, pp. 1–4.
- [39] J. Tang, X. Shu, G.-J. Qi, Z. Li, M. Wang, S. Yan, and R. Jain, "Tri-clustered tensor completion for social-aware image tag refinement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1662–1674, Aug. 2017.
- [40] K. T. Ahmed, S. A. H. Naqvi, A. Rehman, and T. Saba, "Convolution, approximation and spatial information based object and color signatures for content based image retrieval," in *Proc. Int. Conf. Comput. Inf. Sci. (ICCIS)*, Apr. 2019, pp. 1–6.
- [41] A. Khan, "Improved multi-lingual sentiment analysis and recognition using deep learning," *J. Inf. Sci.*, Jan. 2023.
- [42] L. Shui, W. Liu, and Z. Feng, "Automatic image annotation based on generative adversarial network," *J. Comput. Appl.*, vol. 39, no. 7, p. 2129, 2019.



MYASAR MUNDHER ADNAN received the B.E. degree in computer science from Alkufa University, Iraq, in 2011, and the M.S. degree in computer science from UTM, in 2014, where he is currently pursuing the Ph.D. degree. His research interests include deep learning, image processing, and machine learning.



WALEED HADI MADHLOOM KURDI received the master's degree in computer engineering from MMU, India. His research interests include deep learning, image processing, and machine learning.



SAEED ALI OMER BAHAJ received the Ph.D. degree from Pune University, India, in 2006. He is currently an Associate Professor with the Computer Engineering Department, Hadramout University, Yemen, and the MIS Department, COBA, Prince Sattam bin Abdulaziz University. His primary research interests include artificial intelligence, information management, forecasting, information engineering, big data, and information security.

SARAH ALOTAIBI received the B.Sc. and M.Sc. degrees in computer science from King Saud University, Riyadh, Saudi Arabia, and the Ph.D. degree in computer vision from the University of York, U.K. She is currently an Assistant Professor with the Department of Computer Science, King Saud University. Her research interests include computer vision and machine learning, more specifically: statistical modeling, appearance modeling, face modeling, reflectance analysis, inverse rendering using optimization schemes, and deep learning.



MOHAMMED HASAN ALI received the bachelor's degree in computer technical engineering from Islamic University, the master's degree in computer engineering from Universiti Teknikal Malaysia Melaka (UTeM), and the Ph.D. degree in computer science from Universiti Malaysia Pahang (UMP), Malaysia. He is currently a Lecturer with Imam Ja'afar Al-Sadiq University. His research interests include expert systems, machine learning applications, and security.



AMJAD REHMAN (Senior Member, IEEE) received the Ph.D. degree from the Faculty of Computing, Universiti Teknologi Malaysia, specializing in forensic documents analysis and security, in 2010. Currently, he holds the position of a Senior Researcher with the Artificial Intelligence and Data Analytics Laboratory, Prince Sultan University, Riyadh, Saudi Arabia. He was a Principal Investigator (PI) in several funded projects and has successfully completed projects funded by

MOHE, Malaysia, and Saudi Arabia. With over 200 ISI journal articles and conference publications, his H-index stands at 40, with 4000 citations. His research interests include data mining, health informatics, and pattern recognition. Additionally, he was honored with the Rector Award for the Best Student from Prince Sultan University, in 2010.

TANZILA SABA (Senior Member, IEEE) received the Ph.D. degree in document information security and management from the Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia, in 2012. She is currently a Full Professor with the College of Computer and Information Sciences, Prince Sultan University (PSU), Riyadh, Saudi Arabia, and also the Leader of the AIDA Laboratory. She has published over 300 publications in high-ranked journals. Her primary research interests include bioinformatics, data mining, and classification using AI models. She received the Best Student Award from the Faculty of Computing, UTM, in 2012. She received the Best Research of the Year Award from PSU, from 2013 to 2016. Due to her excellent research achievement, she is included in Marquis Who's Who (S&T), in 2012. She is an editor of several reputed journals and on a panel of TPC member of international conferences.

...