

RESEARCH ARTICLE

Visual attention mechanism and support vector machine based automatic image annotation

Zhangang Hao¹*, Hongwei Ge², Long Wang²

1 School of Business Administration, Shandong Technology and Business University, Yantai, Shandong, China, **2** School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning, China

* These authors contributed equally to this work.

* zghao2000@sina.com



Abstract

Automatic image annotation not only has the efficiency of text-based image retrieval but also achieves the accuracy of content-based image retrieval. Users of annotated images can locate images they want to search by providing keywords. Currently most automatic image annotation algorithms do not consider the relative importance of each region in the image, and some algorithms extract the image features as a whole. This makes it difficult for annotation words to reflect salient versus non-salient areas of the image. Users searching for images are usually only interested in the salient areas. We propose an algorithm that integrates a visual attention mechanism with image annotation. A preprocessing step divides the image into two parts, the salient regions and everything else, and the annotation step places a greater weight on the salient region. When the image is annotated, words relating to the salient region are given first. The support vector machine uses particle swarm optimization to annotate the images automatically. Experimental results show the effectiveness of the proposed algorithm.

OPEN ACCESS

Citation: Hao Z, Ge H, Wang L (2018) Visual attention mechanism and support vector machine based automatic image annotation. PLoS ONE 13(11): e0206971. <https://doi.org/10.1371/journal.pone.0206971>

Editor: Qinghui Zhang, North Shore Long Island Jewish Health System, UNITED STATES

Received: June 13, 2018

Accepted: October 23, 2018

Published: November 6, 2018

Copyright: © 2018 Hao et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All Corel5K image set are available from the database: <https://github.com/watersink/Corel5K>.

Funding: This work was supported by the Shandong Provincial Natural Science Foundation, China, by grant no. ZR2017MG022.

Competing interests: The authors have declared that no competing interests exist.

1. Introduction

With the development of digital and Internet technologies, there are massive numbers of digital images on the Internet. It is also becoming harder and harder for users to find these images accurately and quickly. The dramatic increase in the number of images on the Internet combined with the subjectivity, uncertainty and laboriousness of manual annotation has led to a gradual failure to apply text-based image retrieval (TBIR) to large-scale image retrieval [1]. In comparison, content-based image retrieval (CBIR) increases the amount of effort from users, often requiring them to provide initial images as input for the search. So, CBIR has not been used widely [2–5]. Image retrieval process can be indirectly changed into text retrieval by automatic image annotation [6–9]. That is, users need to provide only the query keywords, with the system returning the images associated with the keywords, which is in line with most users' current habits. Automatic image annotation has become an important research topic in

the field of image retrieval. As long as the image can be accurately annotated, good results can be returned to the user.

Segmenting image is very important for automatic image annotation, greatly influencing the annotation results. Most existing automatic image annotation algorithms do not consider the relative importance of different parts of the image to the user when segmenting the image. Other annotation algorithms do not segment the image at all but extract the whole image as a feature. In this paper, visual attention mechanism is introduced into the annotation process. During pre-processing, the image is divided into salient and non-salient regions. During annotation, salient regions of a image are weighted to give them higher priority. This weighting indirectly enhances the performance of image retrieval (because search keywords often reflect the users' interest only in the salient regions) of the image.

2. Related work

Since this paper uses a support vector machine (SVM) for automatic image annotation, we divided the existing automatic image annotation algorithms into two categories: automatic image annotation using a support vector machines and automatic annotation based on other methods.

Automatic image annotation based on SVM: Cusano et al. proposed a classification system using multi-class support vector machines for automatic image annotation which can be used for large-scale video and image management [10]. Gao et al. proposed a hierarchical boosting algorithm to enhance SVM image classifier training for image annotation [11]. Ommasi et al. proposed SVM-based methods to annotate medical images [12]. Verma proposed a new loss function to make the SVM more robust, solving three problems in annotation: incomplete labels, fuzzy labels, structural overlap [13]. Wu et al. proposed a feature selection method, constructing a multi-class classifier using weighted probabilities and an improved SVM [14]. Yu et al. proposed an automatic image annotation algorithm based on a superpixel bag of words model and an SVM classifier [15]. Majidpour et al. used principal components analysis (PCA) to reduce the number of color features followed by an SVM for classification [16]. Jin et al. proposed a multi-label image automatic labeling framework using an improved multi-kernel learning SVM [17]. Hao et al. proposed an automatic image annotation method based on particle swarm optimization (PSO) and support vector clustering (SVC). They use PSO to optimize the SVC for automatic image annotation [9].

Automatic image annotation based on other methods: For image annotation, Makadia et al. proposed a new baseline technique that uses low-level image features and a simple distance algorithm to find the nearest neighbor of a given image [18]. Weston et al. proposed an automatic annotation algorithm for large-scale image datasets that uses simultaneous learning [19]. Liu et al. proposed a multiview Hessian regularization (mHR) algorithm applied to kernel least squares and SVMs for image annotation [20]. Employing jaccard similarities, Johnson et al. used multiple non-parametric image metadata to identify neighbors of related images followed by a deep neural network to annotate images [21]. Tariq et al. proposed a strategy that performed a tensor analysis of new images to evaluate the context, and then combined the evaluated content and the image content to label images [22]. In order to solve the problem with incomplete labels for many training image datasets, Wu et al. used incomplete label data to train classifiers [23]. Shin et al. proposed a deep learning model to detect disease within medical images [24]. Jing et al. proposed a new multi-label learning method that integrates multi-label dictionary learning with embedding of some of the same tags [25]. Rad et al. proposed a new automatic image annotation method based on non-negative matrix factorization [26]. Liu et al. used kernel logistic regression to annotate network images automatically [27].

Choi et al. annotated images by analyzing images and image-related text [28]. Uricchio et al. proposed a label propagation framework based on kernel canonical correlation analysis [29]. Chang et al. performed rapid image annotation using brain state decoding and visual pattern mining [30].

From the literature review, we can see that although there are many works for automatically annotating images, these methods rarely consider the relative importance of each part of the image to users when segmenting image. Therefore, this paper introduces a visual attention mechanism to solve this problem. We use PSO to optimize an SVM, which can improve the accuracy of image annotation.

3. Histogram-based contrast method

The histogram-based contrast method (HC) uses the image global histogram features to extract salient regions of images, in a bottom-up, data-driven approach. The generally accepted theory is that cortical cells of the human brain are genetically predisposed to respond to high contrast stimulation. Based on this theory, a high-resolution, global saliency map contrast analysis approach is proposed:

1. Based on the global contrast method, large-scale objects will be separated from the surrounding environment. This method outperforms a local contrast method that produces significant values only at or near the edge of the object.
2. Global considerations will assign similar salient values to similar regions, consistently highlighting the entire salient object.
3. The significance of the area will depend mainly on the area around it, with distant areas having less impact.
4. The saliency map should be generated simply and quickly to accommodate large-scale image processing and efficient image classification and retrieval.

An HC-Map based on histogram contrast will be used to assign pixel significant values based on color differences from other pixels, producing an image saliency map with full resolution.

3.1. Pixel salient value

Based on the sensitivity of a visual system to the visual signal contrast in biological vision, a histogram-based contrast method is proposed to define the saliency value of each pixel in the image. Specifically, the saliency value calculation for each pixel is obtained by calculating its contrast with every other pixel, that is, the salient value of the pixel in the image is defined as formula (1).

$$S(I_k) = \sum_{\forall I_i \in I} D(I_k, I_i) \quad (1)$$

where $D(I_k, I_i)$ is the distance of pixel I_k and I_i in the Lab color space. Formula (1) can be expanded to formula (2).

$$S(I_k) = D(I_k, I_1) + D(I_k, I_2) + \dots + D(I_k, I_N) \quad (2)$$

where N is the total number of pixels in the image I .

With these definitions, it is easy to see that pixels with the same color value have the same salient value because the spatial relationship of the pixels is ignored. Then, formula (2) can be

rearranged so that the values having the same color are grouped into one class with the saliency value of each color value obtained as a result.

$$S(I_k) = S(c_l) = \sum_{j=1}^n f_i \bullet D(c_l, c_j) \quad (3)$$

where c_l is the color value of the pixel I_k ; n is the number of colors contained in the image; f_i is the probability of c_j in the image.

3.2. Acceleration based on histogram

The time complexity of calculating the salient value of each pixel in the image by formula (1) is $O(N^2)$, which has very high computational cost for a medium sized image. Eq (3) is equivalent to Eq (2), but its time complexity is reduced to $O(N) + O(n^2)$, which means that if $O(n^2) \leq O(N)$, the time complexity is $O(N)$. Therefore, the reduction in the number of pixel colors becomes key. If the RGB color space is used, the maximum number of possible colors per pixel is 256^3 , far greater than the total number of pixels.

In order to reduce the total number of colors, we can reduce the number of colors per channel to 12 from 256 and the total number of colors can be reduced to $12^3 = 1728$. Colors in actual use are only a small percentage of the total color space, and colors that appear less frequently can be discarded to further reduce the number of colors. By selecting the colors that appear at high frequencies, the number of pixels in these colors is guaranteed to account for 95% of the whole image.

3.3. Smoothing color space

Although quantifying colors and extracting colors with high frequency can effectively improve the computational efficiency, there are some shortcomings in the quantification itself. Some similar colors may be incorrectly quantized to colors with different values. To reduce the effect of this random noise on the calculation of saliency values, the use of a smoothing process for each color will improve saliency values. The specific operation replaces the saliency value of each color with a weighted average value of the saliency values of similar colors (measured in terms of distance in the Lab color space). Pick the $m = n/4$ nearest color to improve the color c saliency value.

$$S'(c) = \frac{1}{(m-1)T} \sum_{i=1}^m (T - D(c, c_i)) S(c_i) \quad (4)$$

where $T = \sum_{i=1}^m D(c, c_i)$ is the sum of the distance of the color c from its nearest neighbor m , and the normalization factor is $(m-1)T$.

4. Automatic annotation based on visual attention mechanism and support vector machine

Visual attention mechanism can segment a image into salient and non-salient regions, which is convenient for people to recognize and retrieve the image. In this part, the visual attention mechanism and support vector machine will be combined to complete the annotation of images. This section will detail the whole process of the method.

4.1. Image pre-processing

First, the HC algorithm is used to process the image, dividing the image into salient regions and non-salient regions (Segmentation process is shown in 3). After the image is divided, the image regions are extracted. The original image is divided into two images according to the threshold value in the divided image. One is the saliency object graph, in which the salient region is preserved and the non-salient region is filled with white space. The other is the non-salient object, in which the non-salient areas are preserved, and the salient areas are filled with white space. Then we use a colored pattern appearance model (CPAM) to extract the features of the two images separately [31].

The CPAM model divides the region of the image to be extracted into 4×4 regions, and then extracts the Chromatic Spatial Pattern Histogram (CSPH) and the Achromatic Spatial Pattern Histogram (ASPH) for each small region. The CSPH and ASPH in all small regions are combined to form a feature vector representing the entire image. In this paper, the CSPH and ASPH are represented by a 64-dimensional eigenvector, and each image is represented by a 128-dimensional eigenvector. Supposing two images represented by the CPAM model are x_1 and x_2 respectively, the distance between two images can be calculated by formula (5).

$$d(x_1, x_2) = \sum_{\forall i} \frac{|ASPH_1(i) - ASPH_2(i)|}{1 + ASPH_1(i) + ASPH_2(i)} + \sum_{\forall j} \frac{|CSPH_1(i) - CSPH_2(i)|}{1 + CSPH_1(i) + CSPH_2(i)} \quad (5)$$

where $|\cdot|$ represents the absolute value; $CSPH(i)$ represents the i th component of the image CSPH vector; $ASPH(j)$ represents the j th component of the image ASPH vector.

After the feature is extracted from both the salient and the non-salient regions of the image, each image is represented by two 128-dimensional eigenvectors, namely a salient region feature vector and a non-salient region feature vector. The two eigenvectors are then merged into one eigenvector.

A detailed explanation will be helpful. Because the significance of salient region and non-salient region is different, that is, salient region is more important than non-salient region, it is necessary to give more weight to salient region feature vector to reflect its importance. The size of the salient areas in the image is usually much smaller than those of the non-salient areas. The number of significant area pixels num_S and the number of non-significant area pixels num_N in the image are respectively calculated. Let the weight of the significant region eigenvector be $u_1 = \frac{1}{num_S}$, and the weight of the non-salient region eigenvector be $u_2 = \frac{1}{num_N}$. Because usually $num_S < num_N$, so $u_1 > u_2$. If the salient area of an image is larger than the non-salient area, the values of u_1 and u_2 are swapped to ensure that the weights corresponding to salient area eigenvectors are always larger than the weights corresponding to non-salient area eigenvectors. Finally, the fused eigenvectors are normalized. Supposing the salient regional eigenvector of the image is x_1 , the non-salient region eigenvectors is x_2 , the weighted average and the normalized eigenvectors are x .

$$x = \frac{u_1 x_1 + u_2 x_2}{u_1 + u_2} \quad (6)$$

When the unknown image is marked, the feature vector x is used to represent the image.

4.2. Image model training and annotating

After obtaining the feature vectors of the salient and non-salient regions of an image, the known image is used to train by SVM. This paper specifically uses the support vector data description (SVDD) algorithm [32]. However, considering the literature [9] (our previous

study), we used PSO to optimize the SVDD to improve the accuracy of the algorithm. Originally used to solve one-class classification problems, the SVDD has been later extended to solve multiple classification problems. So, we named the algorithm PVSVD (P refers to PSO; V refers to Visual Attention Mechanism; SVDD refers to SVDD). SVDD differs from other SVM methods, in that it does not need negative sample data in training data, because it is classified by building a hypersphere that can surround all positive sample points as much as possible. SVDD constructs a hypersphere for each category of data, significantly reducing the complexity of solving problems and increasing computational efficiency.

Mathematically, the SVDD is a training set $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\} \in (R^n \times y)^l$, $x_i \in R^n$, $y_i \in y = \{1, 2, \dots, K\}$, $i = 1, 2, \dots, l$. For each class, a hypersphere that encloses all the data points of the training samples is created, defining the class m hypersphere (a_m, r_m) , where a_m is the hypersphere center and r_m is the hypersphere radius. The goal is a problem solution that includes as many points as possible for all training samples while minimizing the radius of the hypersphere. Similar to an SVM, it also allows the sample points to fall outside the hypersphere, with a relaxation variable introduced to obtain the mathematical model of the SVDD problem.

$$\min r_m^2 + C_m \sum_{i: y_i = m} \xi_i \quad (7)$$

$$s.t. \|x_i - a_m\|^2 \leq r_m^2 + \xi_i \quad (8)$$

$$\xi_i \geq 0, \forall y_i = m \quad (9)$$

where C_m (custom constants) is the penalty factor.

When solving a problem, the original problem is transformed into a dual problem.

$$\max_{\alpha} \sum_{i: y_i = m} \alpha_i K(x_i \cdot x_j) - \sum_{i: y_i = m} \sum_{j: y_j = m} \alpha_i \alpha_j K(x_i \cdot x_j) \quad (10)$$

$$s.t. \sum_{i: y_i = m} \alpha_i = 1, 0 \leq \alpha_i \leq C, \forall i : y_i = m$$

$K(x_i \cdot x_j)$ is a kernel function. In this paper, the kernel function is as follows:

$$K(x_i, x_j) = e^{-\frac{d(x_i, x_j)}{h}} \quad (11)$$

where $d(x_i, x_j)$ is defined by formula (5), and h is the dimension of the image feature vector.

For a point x , its distance from the center of the ball is:

$$f(x) = \|x - a_m\|^2 = K(x, x) - 2 \sum_{j=1}^n \alpha_j K(x, x_j) + \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j) \quad (12)$$

When the point x is located outside the ball, there is deformation:

$$\sum_{j=1}^n \alpha_j K(x, x_j) < \frac{1}{2} (1 + \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j) - r^2) \quad (13)$$

Similarly, when point x is located in the ball boundary or inside, there is deformation:

$$\sum_{j=1}^n \alpha_j K(x, x_j) \geq \frac{1}{2} (1 + \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j) - r^2) \quad (14)$$

The kernel density estimation function is introduced for subsequent computation.

$$p(x) = \sum_{i=1}^n \omega_i \varphi(x, x_i) \quad (15)$$

where $\varphi(x, x_i)$ is a window function and ω_i is the weight, where $\sum_{i=1}^n \omega_i = 1$. In this paper, let

$$\varphi(x, x_i) = \frac{1}{n} e^{-\frac{d(x_i, x_j)}{n}} \quad (16)$$

Finally, we get the probability relationship between the m th image and the unknown image by formulas (13)–(16).

$$p(x|m) = \sum_{i=1}^n \alpha_i K(x, x_i) \quad (17)$$

Formula (17) indicates the probability of the image to be marked for the m th class image.

After training, the support vector and the corresponding multipliers for each type of image are obtained. The probability relationship between the image and each type of image can be calculated by formula (17). $p(w_i | I)$ represents the probability that the annotation word of unlabeled image I is w_i . Derived from the Bayesian formula:

$$p(w_i | I) = \frac{p(w_i) p(I | w_i)}{p(I)} \quad (18)$$

where $p(w_i | I)$ is the probability of w_i , which can be obtained from $p(w_i) = n_i / n$, where n_i represents the number of image categories that contain the word w_i and n representing the total number of image categories; $p(I)$ represents the probability of image I , which is fixed for a given image, so the value is not calculated uniformly in the calculation; $p(I | w_i)$ represents the probability of generating I from w_i . In this paper, because the model m of each type image previously trained using SVDD is related to the corresponding keyword w_i , so $p(I | w_i) = p(x | m_i)$. Thus, we can calculate the probability of each word as the unknown image tagging the word, and then take the first four with the highest probability as the annotation words of the image.

4.3. Relationship between the annotation words

There are also some relationships between the marked words. For example, if there is a keyword "sky", "white clouds" is also likely to appear. The probability relationship between words is defined as follows. The probability relationship between all words is dynamically updated before each image is annotated, and the probability of each word annotation image is calculated to improve the annotation result.

Supposing the image collection is as follows:

$$T = \{I_1, I_2, \dots, I_N\} \quad (19)$$

where $I_i, i = 1, 2, \dots, N$ means that the image collection is divided into N classes. Supposing the

annotation word set is as follows:

$$W = \{w_1, w_2, \dots, w_M\} \quad (20)$$

If $I(w_i)$ represents a collection of keywords w_i , then the relationship between keywords w_i and w_j can be expressed as follows:

$$p(w_i|w_j) = \frac{I(w_i) \cap I(w_j)}{I(w_j)} \quad (21)$$

$$p(w_j|w_i) = \frac{I(w_i) \cap I(w_j)}{I(w_i)} \quad (22)$$

In summary, supposing that a set of annotation words of image I is V , and all annotation word sets are U . If $V = \emptyset$, The probability of annotating image I of all words in set U will be calculated.

$$p(w_i|I) = \frac{p(w_i)p(I|w_i)}{p(I)} \quad (23)$$

We take the words with the largest probability as the first tagging words. If $V \neq \emptyset$, the probability of all the annotation words labeled images I in set $U - V$.

$$p(w_i|I, w_1, w_2, \dots, w_M) = \frac{p(w_i)p(I|w_i) \prod_{t=1}^M p(w_t|w_i)}{p(I)p(w_1, \dots, w_M|I)} \quad (24)$$

Therefore, the annotation of the image is given by formula (23) and (24). The probability between words changes due to the frequency of the labeled words. Each time a new image is added, the probability relationship between words changes due to the frequency of the marked words. Every time the image is marked, the probability between all the words is updated.

4.4. Determining the salient words

The determination of salient words takes place according to the procedure given here. After deriving the global annotation words $w_i (i = 1, 2, 3, 4)$ of the image, the following steps are performed for each word: First, find all the image categories M_{ik} containing the word. Then initialize all the distance $d_{ik} = 0$ and calculate the distance d_{ik} between the salient region eigenvector of the image to be marked and the salient model of image category M_{ik} . At the conclusion, only the d_{ik} values corresponding to the words which are included in the marked image and belong to the salient region are greater than 0. The smallest d_{ik} is chosen from all the distance values, and the probability of the annotation word as a salient word for the image is calculated. Finally, the two words with the highest probability are used as salient words for the image, with the remaining two used as non-obvious words. All are output in descending order of probability.

4.5. Algorithm process

The algorithm process follows here steps:

1. Divide the annotated images in the image library and extract the salient areas and non-salient areas.

2. Extract the features from the salient regions and the non-salient regions separately to obtain the corresponding eigenvectors, finally obtaining the eigenvectors that finally represent the whole image by the weighted average of the two.
3. Use the PSO to optimize the SVDD, and the optimized SVDD to train the whole image and salient region eigenvectors. The whole image annotation model and the salient region annotation model are obtained respectively.
4. Place the weighted eigenvector of the image to be annotated into the whole image annotation model image and calculate the overall annotation of the image by combining the relationships between words.
5. Combine the annotation from the analysis of the salient area of the image and the words from the analysis of the whole image and output them firstly. The remaining annotation words are output as non-salient words after the salient words.

This last processing step is possible for two reasons. First, the pre-processing stage uses the visual attention mechanism to divide the image region into salient and non-salient regions with weighted feature vectors representing the image. Second, the training model stage distinguishes the salient and non-salient regions again. Therefore, in the final annotation result, the annotation words of the two parts can be distinguished, with the annotation words corresponding to the salient regions output first to improve image annotation result.

In image retrieval, user searches usually reflect an interest in the salient areas of the corresponding images. If an image annotation method that does not distinguish tag importance is applied to the image retrieval, the system returns all images with tagging words matching the user query, even if the tagging words are of low importance for the image. For example, if a user gives a keyword airplane, some of the images returned by the system to the user are images including an airplane. However, some images may contain an airplane in the picture background, while the user's intention is to find airplanes as the main subject of the image. If the algorithm proposed in this paper is applied to image retrieval, the above problem can be solved, so as to improve the user search experience.

In order to show the training and tagging process in detail, we give Fig 1(training process) and Fig 2(annotation process).

5. Experimental results

5.1. Experimental data sets and experimental environment

Our experiment used the Corel-5K image set [32] containing 5000 images divided into 50 categories and with 10 similar images in each category. In the experiment, 3 to 6 the initial tagging words were provided for each category, with a total number of 63 tagging words. The programming environment was MATLAB. The computer configuration is as follows: 2.60 GHz with 2.0GB RAM. A Gaussian kernel function was used in the SVDD algorithm, with $\sigma = 1 / N$, and penalty factor is $C = 1 / N$. We used $N = 128$ as the dimension of the image feature vector.

5.2. Performance comparison with other algorithms

To evaluate the performance of the proposed algorithm (named PVSVD), the results were measured by recall and precision. We compared PVSVD with five other related algorithms, SVM, KSVM-VT[13], HBF[11], SIA[28] and FICE[22]. We compared the recall and precision of the six algorithms for 63 annotation words. The last result reflects the average recall and precision of all the annotated words. The results are shown in Figs 3 and 4. Although the

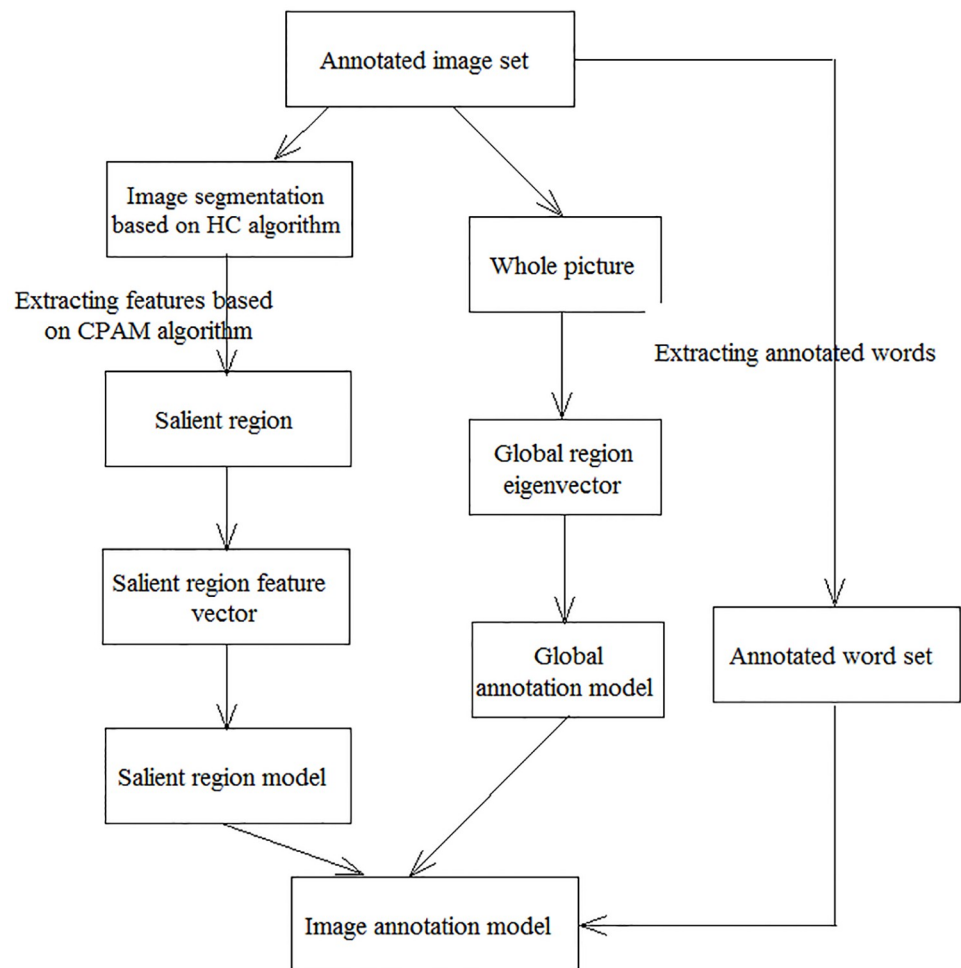


Fig 1. Training process.

<https://doi.org/10.1371/journal.pone.0206971.g001>

performance with some individual words is poor, the PVSVD algorithm is better than the other five algorithms in general.

5.3. Image annotation

Fig 5 shows the annotation results of several test images based on SVM and PVSVD. Each image had four annotation words. The first two are salient words.

When giving image annotation words, the algorithm considers the relationship between words, as described in section 4.3. Before marking each image, the probability relationship between words is updated. For example, if an image is labeled sun, sky, river, clouds, the frequencies of these four words is increased by one, changing the probabilities associated with these four words. As the image library is continuously updated, the probability of all words and other words is also continuously updated. The final image annotation results are affected as a result. Fig 6 shows the comparison of using and not using relations of words. The first two are salient words.

In this paper, the visual attention mechanism is fused into the image annotation process, so that the salient areas in the image can be labeled out first. This is often ignored in past

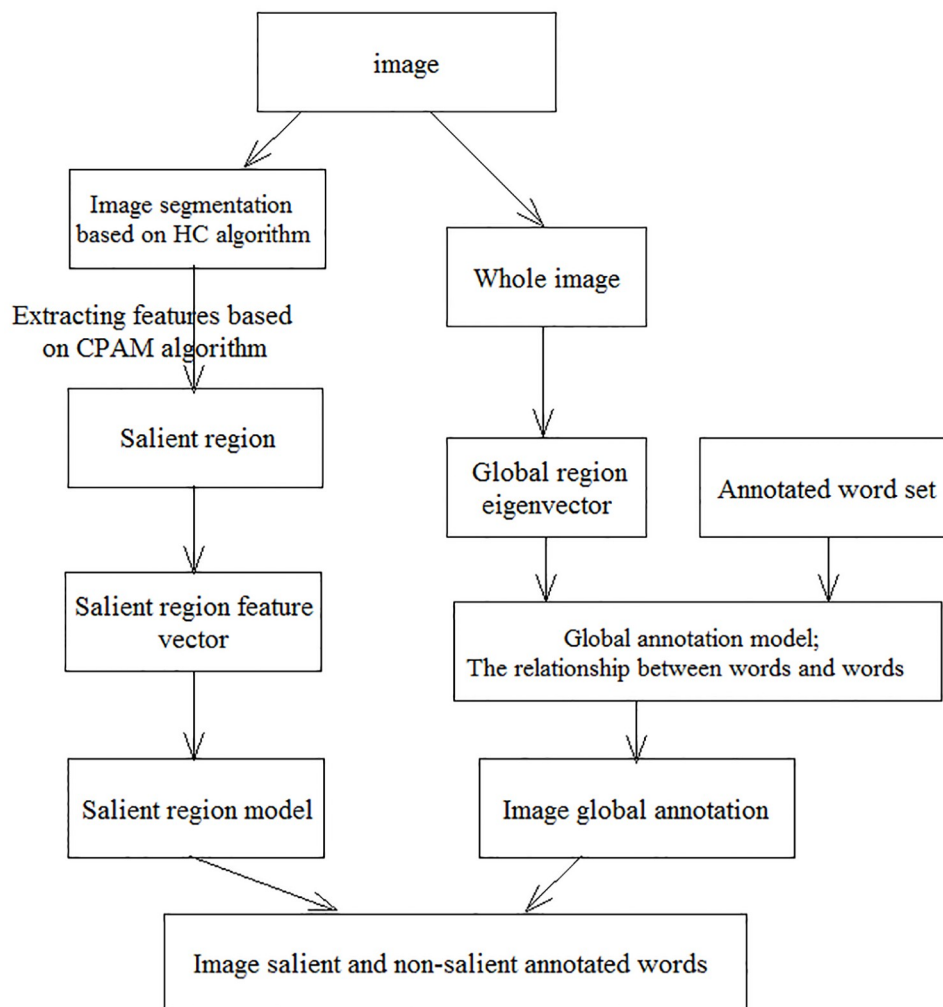


Fig 2. Annotation process.

<https://doi.org/10.1371/journal.pone.0206971.g002>

annotations. As can be seen from the above two tables, the results of annotation using the PVSVD algorithm show the salient areas of the image very well. The annotated words in front are marked the salient regions of the image. These salient regions are the contents that best reflect the content of the image, and are also needed for people to search. The final result can not only reflect the content of the image more accurately, but also improve the retrieval performance of the image.

6. Conclusion

Automatic image annotation not only has the efficiency of text-based image retrieval but also achieves the accuracy of content-based image retrieval. With it, users can find desired images by using keywords. However, until now, almost all the image annotation algorithms extract image features by extracting them without distinguishing the importance of different regions of an image, using the undifferentiated image as training for the annotation process. However, since the keyword given by the user is intended to represent a particular region of interest within the image, if the region of interest of the image can be extracted before annotation, the annotation results can be effectively improved. This leads to better image retrieval.

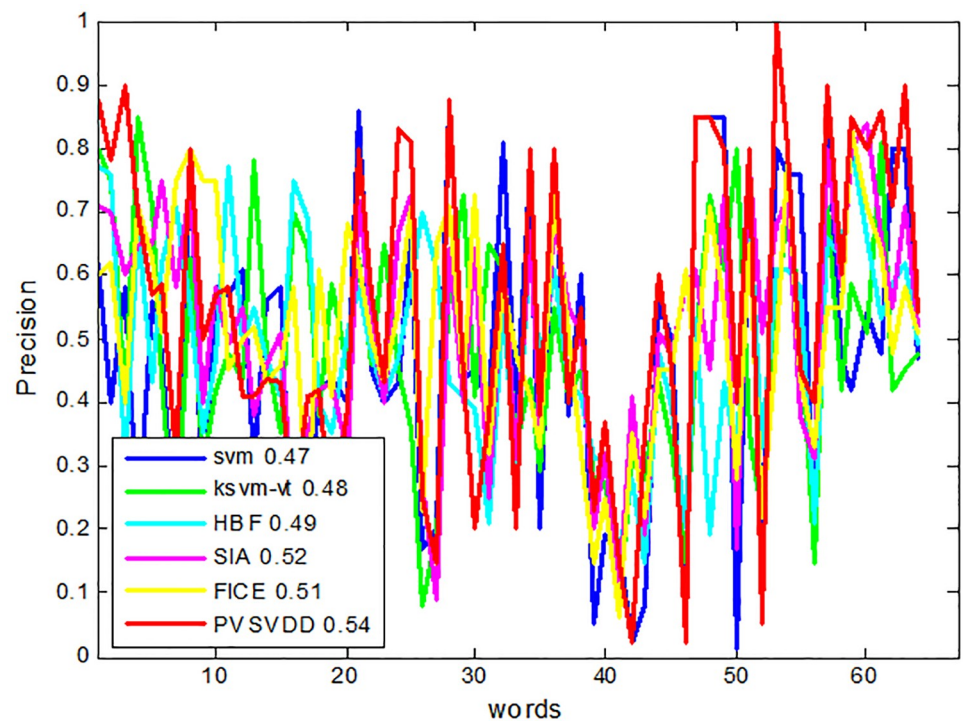


Fig 3. Comparison of precision of six algorithms.

<https://doi.org/10.1371/journal.pone.0206971.g003>

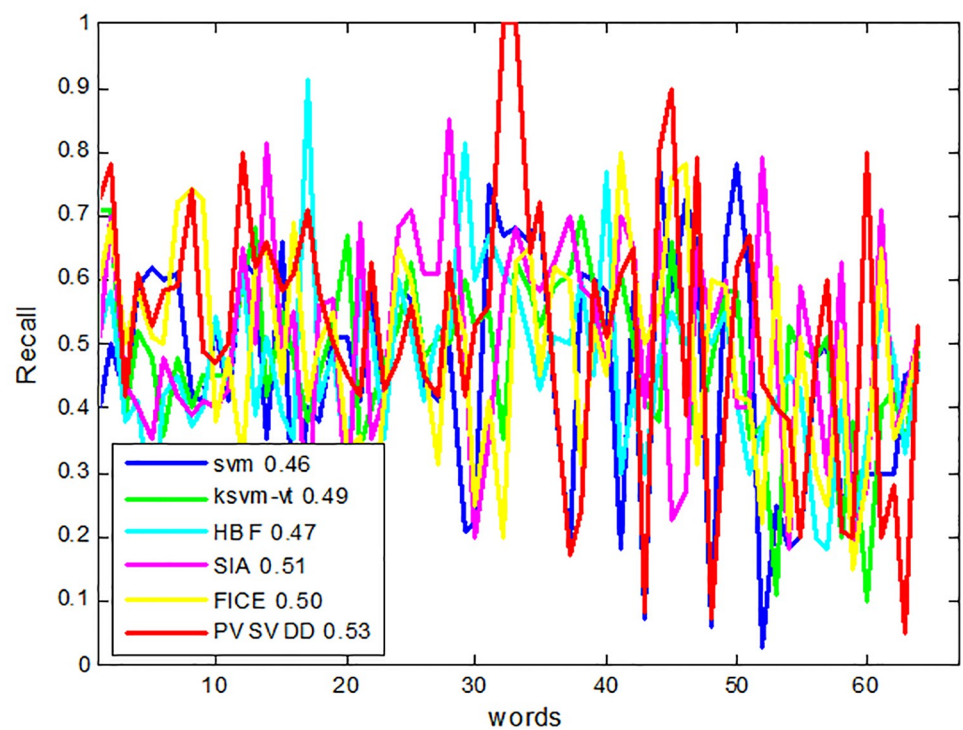


Fig 4. Comparison of recall of six algorithms.

<https://doi.org/10.1371/journal.pone.0206971.g004>

Images			
Annotation	cup	clouds	mountain
result based on	chair	snow	stone
SVM	computer	river	flower
	table	tree	tree
Annotation	computer	river	flower
result based on	table	tree	mountain
PVSVD (The	cup	clouds	tree
first two are	chair	snow	stone
salient words)			
Images			
Annotation	tree	house	jet
result based on	grass	tree	helicopter
SVM	bird	flower	river
	stone	grass	house
Annotation	bird	flower	river
result based on	grass	house	jet
PVSVD (The	tree	tree	helicopter
first two are	stone	grass	house
salient words)			

Fig 5. Example of annotation results. Reprinted from Qian Song under a CC BY license, with permission from Qian Song, original copyright Qian Song 2017.

<https://doi.org/10.1371/journal.pone.0206971.g005>

This paper explores these problems. Image preprocessing extracts salient regions of the image with the addition of the visual attention mechanism. A CPAM algorithm performs feature extraction on salient and non-salient regions, producing the corresponding eigenvectors. After weighting the eigenvectors of the salient regions of the image, the non-salient region eigenvectors are fused to obtain the final eigenvectors of the whole image. The SVDD algorithm based on PSO optimization is used to train and corresponding annotation models are obtained. Finally, the algorithm combines marked words and non-prominent marked words to produce the annotation results.


Images			
Annotation	clouds	chair	house
words that do	grass	house	tree
not use the	tree	tree	grass
relationship	sky	clouds	sky
between words			
Annotation	tree	tree	house
words of PVS	grass	chair	tree
-VDD	clouds	house	grass
	sky	clouds	sky
Images			
Annotation	tree	grass	car
words that do	river	mountain	tree
not use the	stone	tree	house
relationship	clouds	doud	sea
between words			
	river	mountain	sea
Annotation	tree	grass	house
words of PVS	clouds	tree	car
-VDD	stone	clouds	grass

Fig 6. Comparison of using and not using relations of words. Reprinted from Qian Song under a CC BY license, with permission from Qian Song, original copyright Qian Song 2017.

<https://doi.org/10.1371/journal.pone.0206971.g006>

Author Contributions

Data curation: Zhangang Hao.

Project administration: Zhangang Hao, Hongwei Ge.

Software: Zhangang Hao.

Supervision: Zhangang Hao.

Writing – original draft: Long Wang.

Writing – review & editing: Zhangang Hao.

References

1. Samet N, Hiçsönmez S, Şener F. Creating image tags for text based image retrieval using additional corpora[C]. Signal Processing and Communication Application Conference. IEEE, 2016: 1321–1324.
2. Tian Q, Sebe N, Lew M S, Huang TS. Content-based image retrieval using wavelet-based salient points [J]. Proceedings of the SPIE, 2017, 91(4):425–436.
3. Bala A, Kaur T. Local texton XOR patterns: A new feature descriptor for content-based image retrieval [J]. Engineering Science & Technology An International Journal, 2016, 19(1):101–112.
4. Xia Z, Wang X, Zhang L, Qin Z, Sun X, Ren K. A Privacy-Preserving and Copy-Deterrence Content-Based Image Retrieval Scheme in Cloud Computing[J]. IEEE Transactions on Information Forensics & Security, 2017, 11(11):2594–2608.
5. Yu J, Tian Q. Adaptive Discriminant Projection for Content-based Image Retrieval[C]. International Conference on Pattern Recognition. IEEE, 2017:165–168.
6. Uricchio T, Ballan L, Seidenari L, Bimbo AD. Automatic Image Annotation via Label Transfer in the Semantic Space[J]. Pattern Recognition, 2016:1–15.
7. Jin C, Jin S W. Image Distance Metric Learning Based on Neighborhood Sets for Automatic Image Annotation[J]. Journal of Visual Communication & Image Representation, 2016, 34(C): 167–175.
8. Choi D, Kim P. Automatic Image Annotation Using Semantic Text Analysis[J]. 2017, 7465: 479–487.
9. Hao Z, Ge H, Gu T. Automatic Image Annotation Based on Particle Swarm Optimization and Support Vector Clustering[J]. Mathematical Problems in Engineering, 2017(1):1–11.
10. Cusano C, Ciocca G, Schettini R. Image annotation using SVM[C]. Electronic Imaging. International Society for Optics and Photonics, 2003:330–338.
11. Gao Y, Fan J, Jain R, Xue X. Automatic image annotation by incorporating feature hierarchy and boosting to scale up SVM classifiers[C]. ACM International Conference on Multimedia. ACM, 2006:901–910.
12. Tommasi T, Orabona F, Caputo B. CLEF2008 image annotation task: An SVM confidence-based approach[J]. 2008.
13. Verma Y, Jawahar C V. Exploring SVM for Image Annotation in Presence of Confusing Labels[C]. British Machine Vision Conference. 2013:25.1–25.11.
14. Wu W, Nie J, Gao G. An improved SVM-based multiple features fusion method for image annotation[J]. Journal of Information & Computational Science, 2014, 11(14):4987–4997.
15. Min YU, Feng Qin YU, Chen Y. Image annotation based on super-pixel bag of words model and SVM classification[J]. Transducer & Microsystem Technologies, 2016.
16. Majidpour J, Khezri E, Hassanzade H, Mohammed KS. Interactive tool to improve the automatic image annotation using MPEG-7 and multi-class SVM[C]. Information and Knowledge Technology. IEEE, 2015:1–7.
17. Jin C, Jin S W. A multi-label image annotation scheme based on improved SVM multiple kernel learning [C]. Eighth International Conference on Graphic and Image Processing. 2017: 1022510.
18. Makadia A, Pavlovic V, Kumar S. A New Baseline for Image Annotation[C]. European Conference on Computer Vision. Springer-Verlag, 2008:316–329.
19. Weston J, Bengio S, Usunier N. WSABIE: scaling up to large vocabulary image annotation[C]. International Joint Conference on Artificial Intelligence. AAAI Press, 2011:2764–2770.
20. Liu W, Tao D. Multiview Hessian regularization for image annotation [J]. IEEE Transactions on Image Processing, 2013, 22(7):2676. <https://doi.org/10.1109/TIP.2013.2255302> PMID: 23549893
21. Johnson J, Ballan L, Li F F. Love Thy Neighbors: Image Annotation by Exploiting Image Metadata[C]. IEEE International Conference on Computer Vision. IEEE, 2015:4624–4632.

22. Tariq A, Foroosh H. Feature-independent context estimation for automatic image annotation [C]. Computer Vision and Pattern Recognition. IEEE, 2015:1958–1965.
23. Wu B, Lyu S, Hu B G, Ji Q. Multi-label learning with missing labels for image annotation and facial action unit recognition[J]. Pattern Recognition, 2015, 48(7):2279–2289.
24. Shin HooChang, Roberts K, Lu L, Demner-Fushman D, Jianhua Y, Summers RM. Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation[J]. 2016:2497–2506.
25. Jing X Y, Wu F, Li Z, Hu R, Zhang D. Multi-Label Dictionary Learning for Image Annotation[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2016, 25(6):2712–2725. <https://doi.org/10.1109/TIP.2016.2549459> PMID: 27046900
26. Rad R, Jamzad M. Image annotation using multi-view non-negative matrix factorization with different number of basis vectors [J]. Journal of Visual Communication & Image Representation, 2017, 46:1–12.
27. Liu W, Liu H, Tao D, Wang Y, Lu K. Manifold regularized kernel logistic regression for web image annotation[J]. Neurocomputing, 2016, 172(C):3–8.
28. Choi D, Kim P. Automatic Image Annotation Using Semantic Text Analysis[J]. 2017, 7465: 479–487.
29. Uricchio T, Ballan L, Seidenari L, Bimbo AD. Automatic Image Annotation via Label Transfer in the Semantic Space[J]. Pattern Recognition, 2017, 71:144–157.
30. Chang S F, Wang J, Sajda P, Pohlmeyer, Hanna B, Jangraw D. Rapid image annotation via brain state decoding and visual pattern mining[J]. 2017.
31. Sze K W, Lam K M, Qiu G. Scene cut detection using the colored pattern appearance model[J]. 2003, 2:1017–1020.
32. Tax D M J, Duin R P W. Support Vector Data Description[J]. Machine Learning, 2004, 54(1):45–66.

Copyright of PLoS ONE is the property of Public Library of Science and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.