Survey and Analysis on Machine Learning Approaches for Image Annotation

D.O.I- 10.51201/12465

https://doi.org/10.51201/Jusst12465

[1] V. R. Palekar, [2]Dr. Satish Kumar L

[1] Research scholar, SCSE, VIT, Bhopal University, Bhopal, Madhya Pradesh (India)

[1]Assistant Professor, DMIETR, Wardha, Maharashtra (India)

[2]Assistant Professor, VIT, Bhopal University, Bhopal (India)

Abstract: In current years, a large amount of image data is being collected worldwide, which is majorly generated by corporate organizations, health industry and social networking sites. With the strength of substantial level depiction of images, Annotating image has numerous applications not only in image understanding and analysis but also in some of the concern domain like medical research, rural and urban management. Automatic Image Annotation (AIA) has been raised since the late 1990s due to inherent weaknesses of manual image annotation. In this paper, a deep review of the most recent stage in the development of AIA methods is presented by synthesizing 32 literatures published during the past decades. We classify AIA methods into five categories: 1) Kernel Logistic Regression (KLR), 2) Tri-relational Graph (TG), 3) Semantically Regularised CNN- RNN (S-CNN-RNN), 4) Label Correlation guided Deep Multi-view (LCDM), and 5) Multi-Modal Semantic Hash Learning (MMSHL). Considering inspiration on the basis of main idea, framework of model, complexity of computation, time complexity and accuracy in annotation Comparative analysis for various AIA methods are done.

1. Introduction:

Automatic image annotation, which targets to predict the annotation of unknown images according to their relationships between the semantic concept space and the visual feature space, has gained much more attentions in the multimedia research community. Annotation and labelling of Images is highly in demand owing to growth in AI and machine learning (ML) developments. Traditional image annotation technique, where model leaning is done by manual semantic level labelling is not applicable in the exhaustive large scale.

Various types and techniques of image annotation are used to tag images, so that object in image becomes recognizable to machine through computer vision. Many of the researchers are using supervised or completely automatic way of annotating image. With this due research automatic image annotation has achieved makeable gain. support vector machine (SVM) which is supervised learning model uses the classification algorithm and separates the labels dataset in two classes. SVM classifier minimizes the margin between classes using appropriate separating hyperplane. SVM usually reduces the loss caused due to higher margin separating hyperplane classifier. This loss is known as hinge loss which is convex function [2]. Hinge loss is only limited to two classes, it cannot made generalised for multiclass classifier.

On the other hand, supervise approach which deals in manual labelling of images are not applicable for extensive large number of images. One of the solutions is semi-automatic image labelling employing semi-supervised learning (SSL). Generalization ability for limited labelled images can be improved with SSL. SSL explores intrinsic structure from labelled and unlabelled images from training dataset. Manifold regularization is the SSL method which explores the geometry of intrinsic data probability distribution affecting potential and objective function [36, 9, 10]. In last decade the, generative approaches are used to minimize the low-level visual features and high-level semantics gap [23].

2. Related Work:

In last decade most of the development was done regarding multimedia retrieval. Image classification and annotation was developed at semantic and visual level. Mostly the image and video related annotation used learning and model based approach. Normally supervised and semi-supervised methods are used as support vector machine (SVM) and decision tree useful for correlating labels and visual features. looking towards the explosive generation image data, supervised methods are insufficient. Effective labelling is the requirement for generalised learning which was neglected in supervised learning. Considering exhaustive labelling efforts, generalised learning module given by semi-supervised approach reduces the efforts labelling. Generalised learning module is boosted by exploiting less number of labelled data and huge number of unlabelled data. Semi supervised learning attracted the attention at present.

Research Gap/ Objective:

Acquiring meaningful low-level visual features and generating high level semantic correlation is challenging task for Automatic Image Annotation (AIA). Exhaustive large scale image annotation can be done effectively by efficiently indentifying semantic correlating image-image, image-label and label-label

Web images are equipped with additional textual description, this description can be is utilised for effective annotation and efficient retrieval. Usually description is defective which create barrier apply correct annotation method.

Flexible image annotation model is required for large scale worldwide images, to achieve expected prediction.

Research and investigation on reducing semantic gap is highly recommended which requires guidance on neural network choice to train deep computational model for improved efficiency.

3. Methodology

Researchers had developed many machine learning approaches for image annotation in last decade. Most of the methods are belongs to supervised and semi supervised learning.

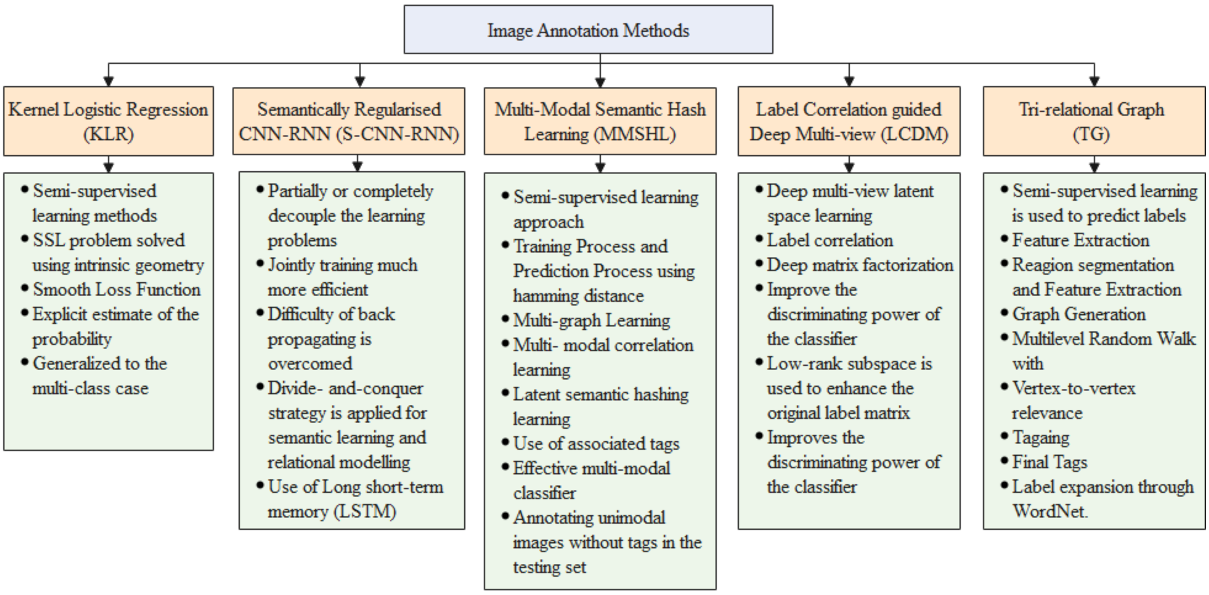
Image Annotation Methods
Kernel Logistic Regression (KLR)
• Semi-supervised learning methods
• SSL problem solved using intrinsic geometry
• Smooth Loss Function
• Explicit estimate of the probability
• Generalized to the multi-class case
Semantically Regularised CNN-RNN (S-CNN-RNN)
• Partially or completely decouple the learning problems
• Jointly training much more efficient
• Difficulty of back propagating is overcomed
• Divide- and-conquer strategy is applied for semantic learning and relational modelling
• Use of Long short-term memory (LSTM)
Multi-Modal Semantic Hash Learning (MMSHL)
• Semi-supervised learning approach
• Training Process and Prediction Process using hamming distance
• Multi-graph Learning
• Multi- modal correlation learning
• Latent semantic hashing learning
• Use of associated tags
• Effective multi-modal classifier
• Annotating unimodal images without tags in the testing set
Label Correlation guided Deep Multi-view (LCDM)
• Deep multi-view latent space learning
• Label correlation
• Deep matrix factorization
• Improve the discriminating power of the classifier
• Low-rank subspace is used to enhance the original label matrix
• Improves the discriminating power of the classifier
Tri-relational Graph (TG)
• Semi-supervised learning is used to predict labels
• Feature Extraction
• Reagion segmentation and Feature Extraction
• Graph Generation
• Multilevel Random Walk with
• Vertex-to-vertex relevance
• Tagaing
• Final Tags
• Label expansion through WordNet.

Fig1. Image Annotation methods

3.1 Manifold regularized Kernel Logistic Regression (KLR) :

Various Machine learning algorithms were developed for Image. Margin between two classes can be maximizing by finding hyperplane employing Support vector machine (SVM). Miximum-margine classifier is being trained by minimizing a hinge loss using SVM. Semi-automatic image annotation is done by applying Semi-supervised learning (SSL) [2,3,4,5].

Executing SVM which uses hyperplane to maximizes the margin between two classes exercises loss. This loss is called as hinge loss. Manifold regularized KLR which is semi supervised approch has the instant advantages first it has a smooth loss function, second it in place of class label it produces an explicit guess of the probability, third it is generalised for multi-class cases; and (4) intrinsic structure of the data distribution can be well utilised by Laplacian regularization [1].

Objective function optimization problem with an additional regularization term to exploit the intrinsic geometry is written as [1, 8, 9, 10].

$$\min_{f \in H_k} \frac{1}{l} \sum_{i=1}^{l} \varphi(f, x_i, y_i) + \lambda_1 \left|\left|f\right|_K^2 + \lambda_2 \right|\left|f_l^2 \qquad (1)$$

Where :

φ - Generalised loss function,

$|f|_K^2$ - K castigate the classifier complexity in proper reproducing kernel Hilbert space (RKHS) H_k,

$|f|_l^2$ - manifold regularization term to castigate f along the underlying manifold,

and λ2 and λ1 balance the regularization terms and loss function $\left|\left|f\right|\right|_l^2$ and $\left|\left|f\right|\right|_K^2$ and respectively.

Local similarity is ensured by laplacian regularization, even thought there are different choices for the manifold regularization terms $\left|\left|f\right|_l^2\right.$

Local similarities are preserved by laplacian regularization. This research as used Laplacian regularized kernel logistic regression to annotate web image [6]. Logistic loss represented with $\log\left(1 + e^{-f}\right)$ used as a loss function to construct a kernel logistic regression[7] (KLR) model. In comparison with supervised SVM, KLR has similar performance.

Therefore, equivalent optimization problem was obtained by incorporating Laplacian regularized term into the objective function with logistic loss.

$$\min_{f \in H_k} \frac{1}{l} \ {}_{i=1}^{l}(y_i \log \frac{1}{1+e^{-f(x_i)}} + \ 1 - y_i)\log(1 - \frac{1}{1+e^{-f(x_i)}}) + \lambda_1 \left|\left|f\right|_K^2 + \lambda_{21} f^T L f\right.$$

[1]

where $f = [f(x_1), f(x_2), \ldots f(x_{l+u})]^T$,

L - graph Laplacian given by L = D - W.

Here D is a diagonal matrix given by $D_{ii} = \sum_{i=1}^{l+u} W_{ij}$,

where W - the edge weight matrix for data adjacency graph[1].

3.2 Semantically Regularised CNN- RNN (S-CNN-RNN),

Various previous studies image annotation techniques have make use of normal CNN-RNN for image annotation even for multilevel classification. Semantic hidden layer of CNN based models developed earlier was not effective. RNN model was overburden with prediction of visual feature and exploring there relations for structure annotation generation, this makes this model slow. It is difficult to train CNN with back propagation through RNN [12].

To reduce the burden from CNN-RNN model, CNN and RNN are separated and hidden layer was added. This hidden layer is semantically trained and imbedded in CNN-RNN, which make

CNN and RNN training separately and parallel. CNN model is semantically trained by taking input image and associated addition information and semantic probability is estimated. Relational modelling is done with estimated probability and correlation model was trained to get sequences of label and words. Concept prediction layer of an Inception net has been used for label prediction in CNN feature layer which is trained in supervision of ground-truth labels/visual concepts, it's clear semantic meaning: Each unit corresponds to a semantic concept. [1, 13].

3.2.1 CNN-RNN:

It is merely important to understand the working of Convolution Neural Network – Recurrent Neural Network (CNN-RNN) before making its user for semantic regularization. A CNN-RNN method is divided as encoding and decoding. Encoder embeds the recognised the visual features of an image and based on embedded features as input decoder generates the sequences of tags and labels.

Image Embedding is represented byI_e . It is a fixed length vector $I_e \in \mathbb{R}^{d \times 1}$. Image Encoder is represented by f_{enc} . Embedding function is represented by $I_e = f_{enc}(I)$. as encoding recognises the visual features and embeds it to image, I_e may be treated as feature transformation[14, 15, 16, 17]. In this method semantic representation is enforced to interact with RNN as it is used as decoder.

Embedded feature from image (I_e) will be passed as context to decoder RNN and predictive path will be generated. Multi label classification may be involved during decoding. Predictive path $\pi = (a_1, a_2, \dots, a_{n_s})$. while generating predictive path importance is given to the sequence of labels in case of multi label classification. n_s represents the number of semantic labels (a_i) predicted for image. During image captioning a_i is the token works from sentence with length a_i. Labels are to be converted in sequence by defining the priority to encounter label imbalance problem. Mostly LSTM-RNN decoder is used as decoder with various CNN as encoder. Training of previous RNN model was affected by messages supports to ups and down of gradient problem. Long Short-Term Remory (LSTM) is widely because of such message controlling mechanism. Cell and hidden are the two states represented by *c* and *h* respectively [11, 19].

Following [11, 18], LSTM-RNN decoder a forward pass at time *t* with input x_t is computed as follows.

Input gate $i_t = \sigma(W_{i,h} \cdot h_{t-1} + W_{i,c} \cdot c_{t-1} + W_{i,x} \cdot x_t + b_i)$

Forget gate $f_t = \sigma(W_{f,h} \cdot h_{t-1} + W_{f,c} \cdot c_{t-1} + W_{f,x} \cdot x_t + b_f)$

Output gate $o_t = \sigma(W_{o,h} \cdot h_{t-1} + W_{o,c} \cdot c_{t-1} + W_{o,x} \cdot x_t + b_o)$ -

Output activation $g_t = \delta(W_{g,h} \cdot h_{t-1} + W_{g,c} \cdot c_{t-1} + W_{g,x} \cdot x_t + b_g)$

Cell state $c_t = f_t \odot c_{t-1} \odot i_t \odot g_t$ - e

Hidden state $h_t = o_t \odot \delta(c_t)$

W·,*h*, *W*·,*c* - recurrent weights,

W·,*x* - input weight, and b· are the biases.

σ(·) is the sigmoid function, and δ is the output activation function.

Decoder uses the last prediction a_{t-1} as input and computes a distribution over possible outputs at time step *t*:

$x_t = E \cdot a_{t-1}$

$h_t = LSTM(x_t, h_{t-1}, c_{t-1}),$

$y_t = softmax(W \cdot h_t + b),$

where *E* - word embedding matrix,

h_{t-1} - hidden state of the recurrent units at $t - 1$,

W, *b* - weight and bias of the output layer,

a_{t-1} - one-hot coding of last prediction a_{t-1}, and LSTM$(\cdot)$ is a forward step of the unit.
The output y_t defines a distribution over possible actions, from which the next action a_{t-1} is sampled.

3.2.2 Semantically regularised CNN-RNN

Likely CNN-RNN as divided the task in two parts, semantically regularised CNN-RNN has reduced the operational load of RNN by dividing it into semantic concept learning and relational modelling. CNN model takes image and associated information as input and generates estimated probabilistic semantic concept. RNN generates the relational model which taken in the estimated probability concept and establish the correlations for generating the sequence of label/words. CNN label concept prediction layer of an Inception net [20] was used instead of feature embedding layer I_e. This embedding has clear semantic concept as it is being trained with ground-truth labels/visual concepts.

For predicting the semantic concept, CNN has huge label space. For multi-label classification approximately 1 k size of label are available. Semantic concepts are predicted these label space $\hat{s} \in \mathscr{R}$^$(k \times 1)$. k is the number of semantic concepts. k is the number of visual concepts are used which normally smaller than the vocabulary word size for captioning the image. RNN generates predictive sequence path π from input$\widehat{S}$. Point to be noted here is that, at both embedding layer $\hat{s}$ and RNN output layer supervision can be added. Which result in concept prediction $\mathscr{L}_u\,(s|\,\hat{s}\,)$ and relational modelling $\mathscr{L}_r\,(\pi, \pi^*\,|\,\hat{s}\,)$ loss.
Formally, we have $\mathscr{L} = \mathscr{L}_u\,(s|\,\hat{s}\,) + \mathscr{L}_r\,(\pi, \pi^*\,|\,\hat{s}\,)$

3.3 Multi-Modal Semantic Hash Learning (MMSHL).

This method uses semi supervised machine learning approach for image annotation. MMSHL model is trained by using labelled and unlabelled image dataset. Researcher has used NUS-WIDE and MRI flicker dataset for experimental results. MRI flicker dataset which has 12500 training and testing samples, 2500 annotated samples, 12500 testing samples, 38 semantic concepts, 457 textual features and 500 image feature is compared with NUS-WIDE dataset having 161789 training and 107858 testing samples, 32357 annotated samples, 81 semantic concepts, 1000 textual features and 100 image feature.

MMSHL model is effective for classify the labelled and unlabelled pair of image-text from training dataset. The intention of this method to annotate unmoral images without tags in its testing set. Annotation method is divided in two steps. First hash function is learned using MMSHL Model on labelled and unlabelled images. Hash function is uses three inputs Multi-grah, Factorization matrix and multimodal correlation. Second the KNN classifier is trained to annotate the image. As this method is using hash function which has efficient storage and computation capacity, it can be used for larger scale image dataset. Associative use of Labels and tags can achieve good result. Modalities of semantic correlations are preserved by this framework [21].

3.3.1 Multi-graph Learning

In multipath learning waited image graph and test graph is used. Semantic correlation between different modalities is identified and multi-modal hashing framework is constructed. Graph matrix for various methods are prepared first and meagre based on their modal graph matrices. This method gives better performance as compared to traditional early fusion and late fusion techniques [22].
Multi-graph learning function is given as [21] :

$$Min\ Tr\left(F^T\sum_{m=1}^{2}\propto_m^2\ L_mF\right)$$

where F - multi-modal semantic matrix,

$\propto_m$ - the weight of modality m,

$L_m = I - A_m$ is the Laplacian matrix of modality m.

A_m is the graph matrix of modality m, which is constructed based on the anchor graph as follows:

$$A_m = Z_m\Lambda_m^{-1}Z_m^T$$

Where diagonal matrix $\Lambda_m = diag\ (Z_m^T1),$

Z_m is the similarity matrix of modality m, which is computed based on anchors as follows:

$$Z_{ij}^m = \exp(-dist_m\ (x_i - x_i^a)/\sigma)$$

Where $x_l^a|_{j=1}^{N=a}$ is the anchor vector,

Feature distance of modality m is given by $dist_m\ (\cdot)$.

For the image modality and text modality, the Euclidean distance and histogram distance, employed respectively. Multi-graph learning process is speedup by semantic matrix$F = ZU$, where U is the semantic mapping matrix, $Z = [Z_1, Z_2]$. To ease the solution of the objective function transformed by taking consideration constraint $U^TU\ = I$ [21].

$$Min\ Tr\left(\sum_{m=1}^{2}\propto_m^2\ U^TZ^TZ_m\Lambda_m^{-1}Z_m^TZU\ \right)$$

s.t. $U^TU\ = I,\ \ \propto_1+\propto_2 = 1$

3.4 Label Correlation guided Deep Multi-view (LCDM)

In existing multi view annotation method the labelled correlation and diversified complex multi view features are ignored which can found in social platform images. Image annotation can be improved by comprehensive description of images. Researches had exposed to correlation of labels in multi-view images [25]. Various features of images are preserved by capturing additional information in data representation. This method explores the correlation of labels by training low level features from label matrix. Originality of label matrix has improved from low level label subset. This technique reduces the missing and noisy labels [24]. Explored label correlation used for training the classifier. Two similar classes are identified using label correlations which improve distinguishing ability of classifier.

3.4.1 DEEP MULTI-VIEW LATENTSPACE LEARNING

Image with multiple views is always represented with compressed data. The representation of this complex data is called latent space. The object in such images may have similarity which makes it difficult to annotate the object. This deep multi-view latent space learning approach has represented unified multi-view data $\{X_v\}_{v=1}^V$ in deep matrix factorization model. Due to this representation coefficient and4 basis matrices are learned layer by layer. Unified multi-view data of all the views represented by consistent coefficient matrix H [23].

To obtain unified data representation from multi-view data$\{X_v\}_{v=1}^V$, we adopt deep matrix factorization model to learn the basis matrices and coefficient matrices layer by layer, and the unified data representation is obtained by introducing a consistent coefficient matrix H across all the views. The minimization objective function is used to reduce the reconstruction error to encode the intra-view correlations better [23].

The optimization is presented as:
$\min_{H,\propto^v} \sum_{v=1}^{V} (\propto^v)^r \|X^v - Z_1^v Z_2^v \ldots Z_m^v H\|_F^2$ $\quad s.t. \sum_{v=1}^{V} \propto^v = 1,\ \propto^v > 0, H \geq 0$
where Z_i^v - the basis matrix of the ith layer for view v,
m - the number of layers,
$\propto^v$ - the weight parameter to control the importance of the v-th view,
H is the learned deep multi-view latent space.
By solving above expression complementary inter-view information can be preserved as each view shares a common representation H by capturing inter-view relations. Weight parameter $\propto^v$ gives the respective view with accuracy due to less embedding loss.

3.4.2 Label Correlation guided Deep Multi-view image annotation (LCDM) method

Label correlation image annotation depends on labelling accuracy. Noisy and missing labels degrade the quality of image annotation. Labelling accuracy can be improved by identifying the missing and noisy labels. To improve the performance this method mainly focused on two tasks. First, Robust label correlation can compete the missing labels and corrects the noisy labels. Second feature based classifier predicts the correct labels by correlation. Comparative results of label correlation, similarity of class are identified. Class identified for two related labels can provide more concrete features than unrelated labels.

The objective function for image annotation,

$$\min_{S,P} \|Y - YS\|_F^2 + \beta\|S\|_* + \eta\|PH - S^T Y^T\|_F^2 + \lambda Tr(P^T LP) \quad s.t. S \geq 0$$

where the first two terms are to learn a low-rank subspace $S \in \mathbb{R}^{C\times C}$ from label Y. Since S captures the correlations of labels, we adopt the constraint S ≥ 0 to ensure the solution is meaningful. The higher value of S_{ij}, the stronger the correla- tion between two labels. The third term is to predict image labels by linear classifier, and $P \in \mathbb{R}^{C\times k}$ is the classifier parameters. P_i is the i-th column of P, which represents the classifier for label l_i. Label correlation S is used to enhance the original image labels, and $S^T Y^T$ is used as the target to train the classifiers. The last term is a graph regularization constraint that imposed on the classifiers. This research introduced affinity matrix of labels $W = \frac{S+S^T}{2}$, and its graph Laplacian
is L = D − W, where D is the diagonal matrix defined as $D_{ii} = \sum_j W_{ij}$ by using l_i and l_j achieve higher correlation, then the corresponding classifier parameters P_i and P_j become more similar. β, η and λ control the importance [23].

3.5 Tri-relational Graph (TG) :

Researcher has observed during this decade that image understanding regarding semantic labelling explored at peak and achieved great attention [27]. Implementing automization in annotation has used image level semantic concepts rather than region level. Visual feature at Image level has limited discrimination power which has less ability to predict small objects [28]. This identified gap has been fulfilled by representing image semantic concept at region level. Visual feature at region level are described more correctly by annotating at region level. But annotating image at region will lead to new problem that image may have several label due various region. These multiple labels are semantically correlated. This problem was addressed by label classification and refinement. The Tri-relational Graph (TG) is mainly designed for web images because relatively good textual description is available for image[26, 35].

3.5.1 Traditional graph verses Tri-relational Graph learning:

Traditional graph based learning uses the data acquired from images only, which uses semi-supervised learning [33, 34], but image data insufficient in image acquiring process. To overcome this, bidirectional graph (BG) is introduced where relationship multiple labels is explored [29]. Image data are represented with multiple labels which require strong semantic relationship. Semantic ambiguity is the issue when correlated multiple labels are assigned based on image data. Semantic ambiguity can be resolved by identifying the various regions in image where multiple labels assigned, which is the motivation to introduced tri-relational graph.

In tri-relational graph annotation method, image is divided in different region and set of various region T is prepared [26, 31]. Set of semantic labels C and Image set X is prepared. Sets of images, regions and labels are prepared based on similarity. Model is trained with Image set I, Semantic labels C, and Region set T.

Based on data of image, labels and region, respective graphs are prepared. New graph invade from image segmentation and label allocation by connecting these subgraphs.

Importance of TG model is vertex to vertex relevance. Ramdon walk restart algorithm is used to find the relation between image, data and label graph, where visual correlation between images and regions, semantic relationship between multiple labels including relevance of image-to-label, label-to-region and region-to-image[26, 30]. Semi supervised approach is used for prediction. Regions with non-labelled images are inserted in TG to predict regions of unannotated image. Researcher has used WordNet [32] for label expansion with the help of nouns and semantics of additional information with web images.

Tri-regional graph semi supervised image annotation having three steps, first generation of tri graph, second annotating region of image based on additional context analysis, third expanding the label using WordNet [26].

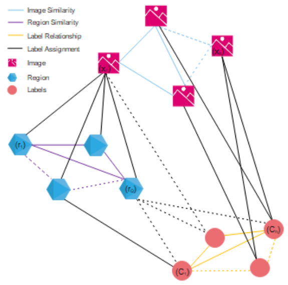

Fig2. Three- Relational Graph [26]

Mainly image is represented in segmented region to extract visual features at low- level. Extracted visual features are analysed and compared then region graph for TG is prepared. To generate Image graph, visual similarity of all the regions from are calculated and compared.

Looking towards the concept the segmentation of the image is important task in TG which is achieved through Texture-enhanced JSEG algorithm which is depend on regional latent semantic dependency [31]. For correct relatively independent segmentation, texture and colour class map are combined by texture-enhanced segmentation (TJSEG). Unnecessary segmentation is more then also performance may penalize, this will happen due to over segmentation. This issue was effectively addressed by point line region (PLR).

SIFT, HSVH, CM and Gabor texture features method are used to represent the region feature and construct the visual word. M_0 X M_0 pixels grid segments are used.

Additional information associated with web image like title, comments and description context is semantically analysed with WorldNet and image contents are described.

As semantic labels are assigned to various region of image union of three properties, image, region and labels are represented as

$$G_q = r_q \cup \{X_i | Y_1(i,q) = 1\} \cup \{c_k | Y_3\,(q,k) = 1\} \qquad - I$$

Here Random Walk Restart [30] (RWR) algorithm which used in birelational graph is modified by using relationship among image, region and labels. Semantics of each group (image, region and labels) are defined as:

$$h_q = \begin{bmatrix} \gamma h_q^x \\ (1-\gamma-\lambda)h_q^R \\ \gamma h_q^L \end{bmatrix} \epsilon \, R_{+}^{n+K+} \qquad \text{- } II$$

Where h_q $(1 \le q \le Q)$

The tri-level random walk expression is formuled as:

$$P_q^{(t+1)}(j) = (1-\propto) \sum p_q^{(t)}\ (i) M\ (i,j) + \propto h_q(i) \quad -III$$

The P_q^* is final distribution which is decided by $P_q^{\infty} = (1-\propto) M^y P_q^{(\infty)} + \alpha\ h_q$ which can be rewritten as:

$$P_q^* = \alpha[\mathrm{I} - (1-\propto)\mathrm{M}^{\mathrm{T}}]^{-1}\mathrm{h_q} \qquad - \mathrm{IV}$$

Tri-level RWR Algorithm for image annotation [26]:

```
Input: Tri-relation Graph: g;
          Transition probability matrix M;
          testing image $X_i$ and its regions $X_i \rightarrow R_i$ , $r_q \in R_i$.
Output: The labels $L_i$ for testing image $X_i$, $c_k \rightarrow r_q$.
1: Insert the test image $X_i$ and its segmented regions $R_i$ into the TG.
2: Analyze the semantic context of the test image according to Equation I.
3: Construct semantic group $H_q$ according to Equation II.
4: Set t = 1, $P_q^t = h_q$
5: repeat 6: Calculate $P_q^{t+1}$ according to Equation III.
7: t= t+1 8: until Equation IV sets up.
9: $c_k = c_{\arg\ max\ p_q^*\ (i)}$
```

4. Contributions of survey method:

1. KLR has solved semi-supervised learning problem using intrinsic geometry. loss function is improvised. Probability is estimated correctly. Multi class cases are generalised.

2. Semantically Regularised CNN-RNN (S-CNN-RNN) has completely decoupled the learning problem. jointly training module is more efficient. Complexity of back propagation is resolved. Semantic learning and relational modelling is done using divide and conquer strategy. Long short term memory (LSTM) was used.

3. Multi-Modal Semantic Hash Learning (MMSHL) has used semi supervised machine learning technique. Hamming distance is used for training and prediction. Model is trained using correlation between multiple graphs and latent semantic hash learning.

Associative tags are used for effective multi-model classifier design having unimodal annotation capability.

4. Label correlation guided deep multi-view (LCDM) has used latent space learning with label correlation. deep matrix factorization is used to improve distinguishing power of classifier. original labels are enhanced by low-rank subspace.

5. Tri-relational graph is semi-supervised learning approach to predict the labels. Features of images and its regions are extracted. Accurate correlation of semantic group and images is done by improving random walk restart algorithm. Image correlation graph is prepared by semantically correlation of image graph, region graph and label graph. WordNet along with additional web information is used for label expansion.

5. Conclusion:

This survey has represented comprehensive study machine leaning approaches AIA methods directed in last decade. Mainly methods are divided into leaning based, training based and model based methods. In this decade researches have moved their direction towards semi supervised learning. Semantic correlation gap of various image representation features is still challenging issue. However exhaustive growth of image data, missing and irrelevant description of image context is the barrier in semi-supervised automatic image annotation which is yet to explore. Approximately 44 zettabytes of data along with images has been generated in 2020. Image recognition from exhaustive large scale image recognition model is facing difficulty of sufficient training images. Identification of unseen classes from without training data will be the future direction.

References

[1] Weifeng Liu, Hongli Liu, Dapeng Tao, Yanjiang Wang, Ke Lu,"Manifold regularized kernel logistic regression for web image annotation," Neurocomputing, Volume 172, 2016, Pages 3-8, ISSN 0925-2312.

[2] Dacheng Tao, Xiaoou Tang, Xuelong Li, Xindong Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval", IEEE Trans. Pattern Anal. Mach. Intell. 28 (7) (2006) 1088–1099.

[3] Weifeng Liu, Dacheng Tao, Multiview hessian "regularization for image annotation", IEEE Trans. Image Process. 22 (2013) 2676–2687.

[4] Weifeng Liu, Dacheng Tao, J.U.N. Cheng, and Yuanyan Tang, "Multiview hessian discriminative sparse coding for image annotation" Comput. Vis. Image Underst., 2013.

[5] Yong Luo, Dacheng Tao, Chang Xu, Chao Xu, Hong Liu, Yonggang Wen, "Multi- view vector-valued manifold regularization for multi-label image classification", IEEE Trans. Neural Netw. Learn. Syst. 24 (5) (2013) 709–722.

[6] Bernhard Schölkopf, Ralf Herbrich, Alex J. Smola, "A generalized representer theorem", Comput. Learn Theory. Lect. Notes Comput. Sci. 2111 (2001) 416–426.

[7] J.I. Zhu, Trevor Hastie, "Kernel Logistic Regression and the Import Vector Machine", J. Computat. Gr. Stat. (2001) 1081–1088

[8] M. Belkin, P. Niyogi, V. Sindhwan i, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples", J. Mach. Learn. Res. 7 (11) (2006) 2399–2434.

[9] N. Guan, D. Tao, Z. Luo, B. Yuan, "Non-negative patch alignment framework", IEEE Trans. Neural Netw. 22 (8) (2011) 1218–1230.

[10] Yong Luo, Dacheng Tao, B.o. Geng, Chao Xu, Stephen J. Maybank, "Manifold regularized multitask learning for semi-supervised multilabel image classifi- cation", IEEE Trans. Image Process. 22 (2) (2013) 523–536

[11] F. Liu, T. Xiang, T. M. Hospedales, W. Yang and C. Sun, "Semantic Regularisation for Recurrent Image Annotation," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 4160-4168, doi: 10.1109/CVPR.2017.443.

[12] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional "random fields as recurrent neural networks". In CVPR, 2015.

[13] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. "NUS-WIDE: A real-world web image database from national university of Singapore". In CIVR, 2009.

[14] J. Jin and H. Nakayama. "Annotation order matters: Recurrent image annotator for arbitrary length image tagging". In ICPR, 2016

[15] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. "Deep captioning with multimodal recurrent neural networks (mrnn)". In ICLR, 2015

[16] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". In ICLR, 2015.

[17] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu. "CNN-RNN: A unified framework for multi-label image clas- sification". In CVPR, 2016.

[18] A. Graves, A. Mohamed, and G. Hinton. "Speech recognition with deep recurrent neural networks". In ICASSP, 2013. 4

[19] S. Hochreiter and J. Schmidhuber. "Long short-term memory". Neural computation, 1997

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. "Rethinking the inception architecture for computer vision". In CVPR, 2016

[21] J. Wang and G. Li, "A Multi-modal Hashing Learning Framework for Automatic Image Annotation," 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC), Shenzhen, 2017, pp. 14-21, doi: 10.1109/DSC.2017.48.

[22] Snoek C G M, Worring M, Smeulders A W M. "Early versus late fusion in semantic video analysis" , ACM International Conference on Multimedia, Singapore, November. DBLP, 2005:399-402.

[23] Z. Xue, J. Du, M. Zuo, G. Li and Q. Huang, "Label Correlation Guided Deep Multi-View Image Annotation," in IEEE Access, vol. 7, pp. 134707-134717, 2019, doi: 10.1109/ACCESS.2019.2941542.

[24] Z. Lin, G. Ding, M. Hu, J. Wang, and X. Ye, ''Image tag completion via image-specific and tag-specific linear sparse reconstructions,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2013, pp. 1618–1625

[25] P. Zhu, Q. Hu, Q. Hu, C. Zhang, and Z. Feng, ''Multi-view label embed- ding,'' Pattern Recognit., vol. 84, pp. 126–135, Dec. 2018.

[26] Jing Zhang, Ti Tao, Yakun Mu, Han Sun, Dongdong Li, Zhe Wang, "Web image annotation based on Tri-relational Graph and semantic context analysis, Engineering Applications of Artificial Intelligence", Volume 81, 2019, Pages 313-322, ISSN 0952-1976

[27] Zhang, D., Islam, M.M., Lu, G., 2012. "A review on automatic image annotation techniques. Pattern Recognit". 45 (1), 346–362.

[28] Zhang, J., Wu, Q., Shen, C., Zhang, J., Lu, J., 2016." Multi-Label Image Classification with Regional Latent Semantic Dependencies", CoRR, abs/1612.01082, arXiv:1612. 01082.

[29] Wang, H., Huang, H., Ding, C., 2011. "Image annotation using bi-relational graph of images and semantic labels". In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, pp. 793–800

[30] Fellbaum, C., Miller, G., 1998. "Wordnet : an electronic lexical database". Libr. Quart. Inf. Community Policy 25 (2), 292–296

[31] Zhang, J., Mu, Y., Feng, S., Li, K., Yuan, Y.-B., Lee, C.-H., 2018. "Image region annotation based on segmentation and semantic correlation analysis". IET Image Process. 12 (8), 1331–1337

[32] DeviantArt, http://www.deviantart.com.

[33] Chen, G., Song, Y., Wang, F., Zhang, C., 2008."Semi-supervised multi-label learning by solving a sylvester equation". In: Proceedings of the 2008 SIAM International Conference on Data Mining. SIAM, pp. 410–419.

[34] Tong, H., Faloutsos, C., Pan, J.Y., 2006. "Fast random walk with restart and its applications". In: International Conference on Data Mining. pp. 613–622.

[35] V. Palekar, M. Ali and R. Meghe, "DEEP WEB DATA EXTRACTION USING WEB-PROGRAMMING-LANGUAGE-INDEPENDENT APPROACH" Journal of Data Mining and Knowledge Discovery ISSN: 2229–6662 & ISSN: 2229–6670, Volume 3, Issue 2, 2012, pp.-69-73

[36] M. Belkin, P. Niyogi, V. Sindhwan i, "Manifold regularization: A geometric framework for learning from labeled and unlabeled example"s, J. Mach. Learn. Res. 7 (11) (2006) 2399–2434.