

Image Annotation in Presence of Noisy Labels

Chandrashekar V¹, Shailesh Kumar², and C V Jawahar¹

¹ IIIT-Hyderabad, India

² Google Inc., Hyderabad

Abstract. Labels associated with social images are valuable source of information for tasks of image annotation, understanding and retrieval. These labels are often found to be noisy, mainly due to the collaborative tagging activities of users. Existing methods on annotation have been developed and verified on noise free labels of images. In this paper, we propose a novel and generic framework that exploits the collective knowledge embedded in noisy label co-occurrence pairs to derive robust annotations. We compare our method with a well-known image annotation algorithm and show its superiority in terms of annotation accuracy on benchmark Corel5K and ESP datasets in presence of noisy labels.

Keywords: *Image Annotation, Semantic concepts, Graph Mining*

1 Introduction and Related Work

Over the years, the Internet has become the largest database for multimedia content and is organized in a rich and complex way through tagging activities. One such example is collaborative tagging websites, such as Flickr, which collects millions of photos per month from tens of thousands of users. Consequently, there is immense research interest in producing efficient image annotation techniques for labelling social images to cope with the continuously growing amount of social image data.

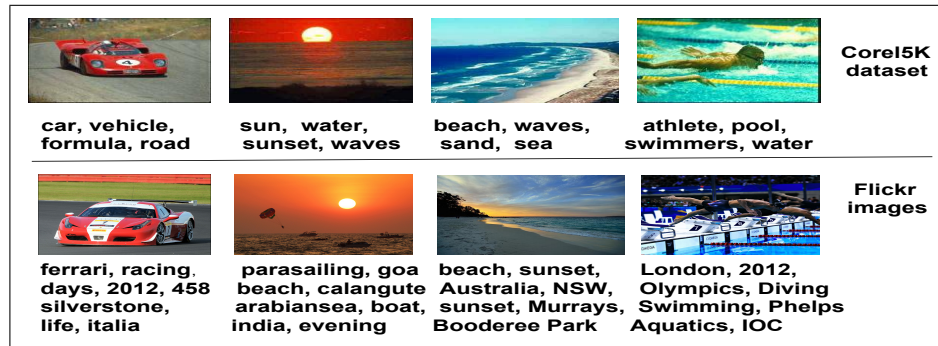


Fig. 1. Figure shows example images and corresponding labels from both Corel5K datasets and Flickr images.

Existing annotation methods [1–4] consider the labels associated with the images to be devoid of errors and belonging to a small fixed vocabulary, and hence, can be directly used for designing annotation schemes. In contrast, the labels collected by collaborative tagging websites are noisy i.e misspelled, redundant, irrelevant to content, and unlimited in numbers. Thus, an interesting problem

to address is, on how to use the noisy information available for annotating unlabelled images reliably.

Figure 1 shows examples of images from both expert-annotated Corel5K dataset and user tagged Flickr images. Labels of Flickr images like *love, life, emotions, excited etc.* are large in number and also irrelevant to the image content, whereas labels of Corel5K are often small in number and precisely describe the content of the image. In this paper, we address the problem of image annotation in presence of noisy labels.

Methods like WSABIE [3], which learn a low dimension embedding space for images and annotations, address this issue in an indirect way. Even Wordnet-based approach [5] has been used to remove irrelevant labels. In MLFDA [6], image annotation is posed as a multi-modal multi-class classification problem, where the noisy data is treated as a special kind of modal of the class and separating hyperplanes between classes are learned by kernel-based LFDA.

In this paper, we address the task of image annotation on noisy data using concept-modelling, a very popular notion in Information Retrieval community. The intuition is that, a specific meaning or aspect of an image can be well described by a group of highly related labels, referred to as label concept. Accordingly, each image can be organized into groups, each of which matches one label concept. This type of image organization not only removes noisy labels associated with an image, but also predicts additional labels that are actually depicted in the image but missing in the ground-truth annotations.

To show the utility of *concepts* over noisy systems, we compare its annotation performance with a baseline annotation method JEC [1], with noisy labels on Corel5K and ESP datasets. Our experimental results suggest that the proposed concept-based method leads to superior image annotation performance compared to JEC in presence of noisy labels.

2 Image Annotation in Presence of Noisy Labels

2.1 Nearest Neighbour Model for Annotation

K-nearest neighbour (or KNN) based methods [1, 2, 4] have been found to give some of the best results on the task of image annotation. The intuition behind them is that similar images share common labels. Most relevant KNN-based annotation methods are (i) JEC [1], which treats the annotation problem as retrieval and proposes a greedy algorithm for label transfer from neighbours, (ii) TagProp [4], a weighted KNN based method that transfers labels by taking weighted average of labels present among the neighbours, and (iii) 2PKNN [2], where a class-wise semantic neighbourhood is defined and only samples within this neighbourhood are used for annotation of unseen image. Since JEC is the essential backend method for modern successful techniques [2, 4], we compare our results with JEC [1] and show that our method is robust under noisy labels.

Let $I = \{ I_1, \dots, I_N \}$ denote the collection of images and $V = \{ v_1, \dots, v_m \}$ denote the *vocabulary* of m labels. The training set $T = \{ (I_1, V_1) \dots (I_N, V_N) \}$ consists of pairs of images and their corresponding label sets, with each $V_i \subseteq V$. Given an unannotated image J , the task of annotation is to predict a set of

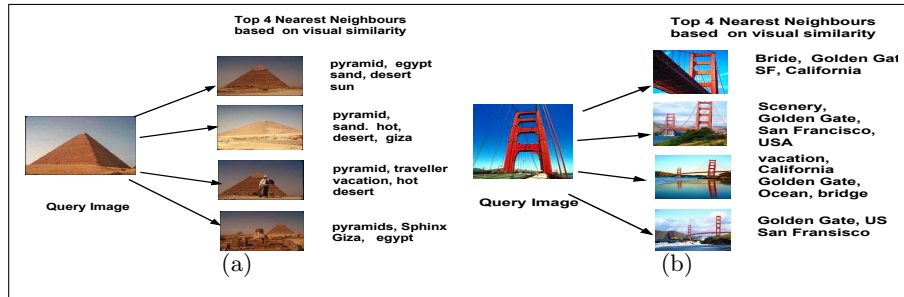


Fig. 2. Figure illustrates example of the well known K-Nearest Neighbour(K-NN) model used in image annotation. For each query image, the top 4 visually similar images are shown, along with the labels associated with them. The labels of the nearest neighbours are transferred to the query image for annotation.

labels that semantically describe J . In a typical NN-setting, we pick the top K visually similar images $T_J = \{ (T_{J,1}, \gamma_{J,1}) \dots (T_{J,K}, \gamma_{J,K}) \}$. $\gamma_{J,K}$ denotes the visual similarity score of image J with its K^{th} neighbour, defined as:

$$\gamma_{J,K} = VisualSimilarity(I_J, T_{J,K}) \quad (1)$$

This score is generated as a function of distance between the images in visual feature space (SIFT, Color Histograms, GIST). Then, the labels of the nearest neighbours are ranked on basis of a label scoring function, κ_{J,v_i} and the top L labels are used to annotate the test image J . This label scoring function is usually based on frequency [1] or distance [4]. Figure 2 shows illustration of KNN model for image annotation.

2.2 Noisy Labels

In photo sharing websites, such as Flickr and Picasa, it is believed that most of the labels are correct, although there are many incorrect and redundant labels. Even from Figure 1, it can be observed that around 40-50% of the labels are irrelevant and are out-of context of the concept which the image represents.

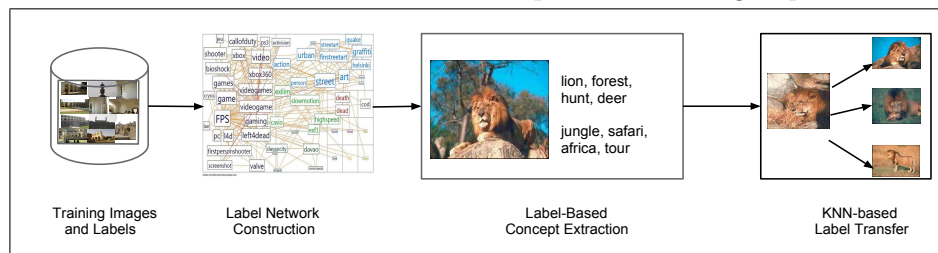


Fig. 3. An overview of our approach, which includes label network construction based on their co-occurrence, semantic concept identification using image labels and a KNN-based approach for transferring labels of concepts to unannotated image.

Existing KNN-based methods [1, 2, 4] make an inherent assumption that labels present in the training set are reliable and correct, and hence can be directly

used for training. They do not have an implicit mechanism of handling noisy labels and would not be suitable for annotation task in collaborative systems.

In this paper, we first present a graph-based approach for exemplifying the relationships between labels along with a noise removal algorithm to remove most of the semantically-unrelated links among the labels. We then make use of this label network to infer the semantic concepts associated with images. Finally we illustrate how these concepts could be used for image annotation in a KNN-based setting. Figure 3 summarizes our approach.

2.3 Label Network Construction and Noise Removal

For label network creation, first the label co-occurrence counts, $\psi(\alpha, \beta)$ $\alpha, \beta \in V$, are calculated. But, this is not the best measure to quantify label co-occurrence strength as it may happen that two very frequent but uncorrelated tags might co-occur a lot compared to two relatively rare but correlated tags. Hence, we use *consistency*, $\phi(\alpha, \beta)$, to quantify associativity between labels, which is loosely defined as how much more likely is it to see the two labels together than random chance. We start by computing three types of raw statistics from the labels of training images: (i) Co-occurrence Counts $\psi(\alpha, \beta)$, (ii) Marginal Counts $\psi(\alpha)$, and (iii) Total Counts ψ_0 (defined in Step 3 of Algorithm 1). Joint probabilities $P(\alpha, \beta) = \frac{\psi(\alpha, \beta)}{\psi_0}$ and marginal probabilities $P(\alpha) = \frac{\psi(\alpha)}{\psi_0}$ are computed from these counts. These statistics are used for computing pair-wise consistencies between labels. We use *Normalized Point-Wise Mutual Information*³, defined as

$$\phi(\alpha, \beta) = \frac{\log\left(\frac{P(\alpha, \beta)}{P(\alpha)P(\beta)}\right)}{-\log P(\alpha, \beta)} \forall \psi(\alpha, \beta) > 0 \quad (2)$$

to exemplify this consistency between labels, as it is a well-bounded quantity and suitably satisfies the definition of consistency.

Algorithm 1 DENOISE($[\psi(\alpha, \beta)]$)

- 1: Iteration $t \leftarrow 0$
 - 2: $\psi^{(t)}(\alpha, \beta) \leftarrow \psi(\alpha, \beta)$
 - 3: $\psi^{(t)}(\alpha) \leftarrow \sum_{\beta \in V} \psi^{(t)}(\alpha, \beta)$, $\psi_0^{(t)} \leftarrow \frac{1}{2} \sum_{\alpha \in V} \sum_{\beta \in V} \psi^{(t)}(\alpha, \beta)$
 - 4: **while** $\sum_{\alpha \in V} \sum_{\beta \in V} \phi^{(t)}(\alpha, \beta)$ converges **do**
 - 5: $\phi^{(t)}(\alpha, \beta) \leftarrow \text{CONSISTENCY}\left(\psi^{(t)}(\alpha, \beta), \psi^{(t)}(\alpha), \psi^{(t)}(\beta), \psi_0^{(t)}\right)$
 - 6: $\psi^{(t+1)}(\alpha, \beta) \leftarrow \psi^{(t)}(\alpha, \beta) \delta\left(\phi^{(t)}(\alpha, \beta) > \theta_{consy}\right)$ $\{\delta(bool) = 1$ if $bool$ is $true$ else $0.\}$
 - 7: $\psi^{(t+1)}(\alpha) \leftarrow \sum_{\beta \in V} \psi^{(t+1)}(\alpha, \beta)$, $\psi_0^{(t+1)} \leftarrow \frac{1}{2} \sum_{\alpha \in V} \sum_{\beta \in V} \psi^{(t+1)}(\alpha, \beta)$
 - 8: $t \leftarrow t + 1$
 - 9: **end while**
-

Iterative Noise Removal: Initially, there is insufficient knowledge to identify which label-pairs are noise. After computing consistencies, label pairs with consistencies lower than a threshold θ_{consy} can be declared noise and are removed from the network. The marginal and total counts are then updated and

³ http://en.wikipedia.org/wiki/Pointwise_mutual_information

consistencies are recomputed in the next iteration. The iterative noise removal method is described in Algorithm 1.

Table 1 shows effect of noise removal on the pair-wise consistencies of labels associated with label *water* in Corel5K data. It can be seen that consistencies of label *water* with correlated labels like *sea*, *ocean*, *beach*, *lake* increases significantly, whereas with irrelevant labels like *hills*, *grass* decreases to zero. By the end of this phase, we obtain a clean **noise-free** label network with pair-wise consistencies between labels as edge weights, which we will call *label consistency network*.

Label	sea	ocean	beach	lake	pool	hills	grass
Before Denoising	0.3257	0.3720	0.3195	0.1699	0.0148	0.2081	0.1461
After Denoising	0.5750	0.5728	0.5658	0.3629	0.2449	0	0

Table 1. Effect of noise removal on the consistencies of labels associated with label **water** in noisy Corel5K dataset, with $\theta_{consy} = 0.01$.

2.4 Label-based Concept Extraction

Here, we use the label consistency network for identifying semantic concepts associated with training images, using the image labels as seed. We define concepts as local maximal subgraphs in the label consistency network, based on a novel measure of *concept strength*, which is in-turn defined in terms of *label strength*.

In a systematic way, we first define *label strength* of a node (label) in a subgraph as a measure that captures the connectivity of the node with rest of the nodes in the subgraph. This essentially is the eigenvector centrality [7] of the subgraph. If $\mathbf{x} = \{x_1, x_2, \dots, x_m\}$ be a set of m nodes in a subgraph and $\mathbf{W}(\mathbf{x}) = [\phi(x_i, x_j)]$ be the label consistency submatrix associated with this subgraph, then by eigenvector centrality, the *label strengths* converge to the first *unnormalized* eigenvector of $\mathbf{W}(\mathbf{x})$.

If $\lambda_1(\mathbf{W}(\mathbf{x}))$ is the first eigenvalue and $\mathbf{v}_1(\mathbf{W}(\mathbf{x}))$ is the first (normalized) eigenvector of this matrix, then *label strengths*, $\rho(\mathbf{x}|\mathbf{W}(\mathbf{x}))$, are defined by:

$$\rho(\mathbf{x}|\mathbf{W}(\mathbf{x})) = \lambda_1(\mathbf{W}(\mathbf{x})) \times \mathbf{v}_1(\mathbf{W}(\mathbf{x})) \quad (3)$$

$$\pi(\mathbf{x}|\Phi) = \min_{i=1 \dots m} \{\rho_i\} \quad (4)$$

To capture the *tightness* of an arbitrary subgraph, we define *concept strength*, $\pi(\mathbf{x}|\Phi)$, to be **minimum** of the *label strengths* of all nodes (labels) of the subgraph (Equation 4). We now define *concepts*, as all those subgraphs in the label network, whose *concept strength* is higher than all its “neighbours”. Neighbours of a subgraph, \mathbf{x} , is defined as all the subgraphs which can be obtained either by adding a single node ($\mathcal{N}_+(\mathbf{x})$) or removing a single node ($\mathcal{N}_-(\mathbf{x})$) from the given subgraph.

$$\mathcal{N}_+(\mathbf{x}) = \{\mathbf{y} = \mathbf{x} \oplus v \mid \forall v \in \mathbf{V} \setminus \mathbf{x}\} \quad \mathcal{N}_-(\mathbf{x}) = \{\mathbf{y} = \mathbf{x} \setminus v \mid \forall v \in \mathbf{x}\} \quad (5)$$

We propose a greedy label-based approach to find such concepts, using the two atomic operations of **grow** and **shrink**. The **grow** operation tries to exhaustively find the best subgraph in $\mathcal{N}_+(\mathbf{x})$, which will have maximum *concept strength*, whereas the **shrink** operation finds the best subgraph in $\mathcal{N}_-(\mathbf{x})$.

Algorithm 2 explains how we extract multiple concepts associated with an image, using their labels as seed. The algorithm iterates over two phases: (i) **Shrink Phase**, which reduces labelset into a candidate subset of highly correlated labels, and (ii) **Grow Phase**, which adds more correlated labels to the candidate set making it a complete concept. The residue of the shrink phase is then again used as input over the next iteration to identify more concepts. Over this recursive process, multiple concepts associated with an image are identified.

Algorithm 2 Label Based Concept Extraction(\mathbf{x}_0, Φ)

```

1:  $\mathbf{x} \leftarrow \mathbf{x}_0$ 
2:  $\mathbf{x}_{conc} = [ ]$ 
3: while  $\mathbf{x}$  do
4:    $[\mathbf{x}_{cand}, \mathbf{x}_{rem}] = \text{ShrinkPhase}(\mathbf{x}|\Phi)$ 
5:    $\mathbf{x}_{conc} = [ \mathbf{x}_{conc} \text{ GrowPhase}(\mathbf{x}_{cand}|\Phi) ]$  {Concepts extracted are concatenated}
6:    $\mathbf{x} \leftarrow \mathbf{x}_{rem}$ 
7: end while
8: A. ShrinkPhase( $\mathbf{x}|\Phi$ )
9: loop
10:   $[\mathbf{x}^-, \mathbf{x}_{rem}] \leftarrow \text{Shrink}(\mathbf{x}|\Phi)$  {Best possible down-neighbor.}
11:  if  $\pi(\mathbf{x}^-) > \pi(\mathbf{x})$  then
12:     $\mathbf{x} \leftarrow \mathbf{x}^-$  {Not reached local maxima yet.}
13:  else
14:    return  $[\mathbf{x}, \mathbf{x}_{rem}]$  {Reached local maxima.}
15:  end if
16: end loop
17: B. GrowPhase( $\mathbf{x}|\Phi$ )
18:  $\mathbf{x}^+ \leftarrow \text{Grow}(\mathbf{x}|\Phi)$  {Best possible up-neighbor.}
19: while  $\pi(\mathbf{x}^+) > \pi(\mathbf{x})$  do
20:    $\mathbf{x} \leftarrow \mathbf{x}^+$ 
21:    $\mathbf{x}^+ \leftarrow \text{Grow}(\mathbf{x}|\Phi)$ 
22: end while
23: return  $\mathbf{x}^+$ 

```

2.5 Label Transfer for Annotation

Now, we illustrate how *concepts* can be used for the task of image annotation algorithm in an NN-setting. As a pre-processing step, we first use the training image labels to create a label consistency network and concepts associated with individual training images are extracted.

Given a test image J , we first find the top K -visually similar training images using features and metrics as suggested in [4]. Then, the labels associated with the concepts of the nearest training images are ranked based on a label scoring function, κ_{J,v_i} , defined as:

$$\kappa_{J,v_i} = \gamma_{J,K} \cdot \pi(\mathbf{x}|\Phi) \cdot \rho_{v_i}(\mathbf{x}) \quad (6)$$

This score is computed as product of visual similarity of the training image to the test image, concept strength of the concept associated with the training image and the label strength of the label within the concept. The individual

components of the scoring function are first normalized before computing the scores. The labels are ranked based on this score and top L unique labels are assigned to the test image. Please note same label could have multiple scores, due to presence of same label in multiple concepts or images.

3 Experiments

We present both qualitative and quantitative results, showing comparisons of our method with a very popular baseline method JEC [1], on benchmark annotation datasets: Corel5K and ESP [1].





			
lion, water, grass, forest	elephant, river, trunk	desert, pyramid, sun, sand,	bay, water, bridge
lion, grass, forest, birds, hills	elephant, river, water, sky, clouds	pyramid, sun, desert, beach, sunrise	water, ocean, sea, bridge, clouds
lion, water, grass, forest, river	elephant, clouds, river, sky, trunk	pyramid, desert, sand, sun, dunes	sea, water, bridge, bay, hills

Fig. 4. Annotation of test images from noisy Corel5K dataset. The second, third and fourth rows show the ground truth labels, the labels predicted by JEC and the labels predicted by our method respectively. The labels in red are those, though depicted in the corresponding images, are missing in the ground-truth annotations and are predicted by our method.

As the motivation of our work is to show the effectiveness of our method on data with noisy labels, we create a parameter modulated noisy dataset by adding noisy labels to the training images of Corel5K and ESP. The noisy labels are taken from a vocabulary which has no overlap with the ground-truth vocabularies. We perform modulated experiments by regulating the degree of noise added to training data, using a parameter Q , which denotes the number of noisy labels added per training image. Annotation models are created using both, our method and JEC [1]. Evaluations are done using popular metric of mean F1-score over all the labels in the original vocabulary of the dataset. The F1-scores reported by our method correspond to label networks with threshold $\theta_{consy} = 0.01$, which was experimentally observed to be giving best results.

Figure 4 shows some qualitative results obtained using our method and JEC on noisy Corel5K data. It can be seen that some labels predicted by JEC are irrelevant and, also some ground truth annotations are missing in the predictions, whereas our method predicts all ground-truth annotations along with labels, which are depicted in the image but missing in the ground-truth annotation.

To analyze our method’s performance quantitatively, we compute the F1-score of each label in the ground-truth. The mean F1 scores using our method as well as those obtained by JEC [1] are reported in Figure 5. In both Corel5K and ESP datasets, as noise increases, F1-score of both methods decrease, but relatively our method performs better than JEC. In Corel5K, when only one

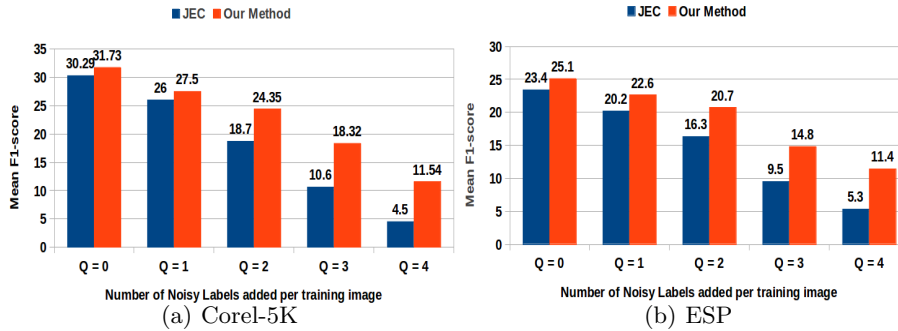


Fig. 5. Comparison of annotation performance of our method and JEC[1] on noisy Corel5K and ESP datasets. Q denotes the number of noisy labels per training image. noisy label is added per training image ($Q = 1$), there is about 6% improvement in F1-score. As Q is increased to 4, there is around 150% increase in the F1-score, which is a very significant improvement. This shows the effectiveness of using *concepts* in the task of image annotation, especially when noise is too high.

Experimentally we found that as θ_{consy} increases, the F1 scores also increase upto to a saturation point, and then start decreasing. This happens because once θ_{consy} reaches its saturation value, even relevant label-pairs in the network are considered as noise and discarded in the noise removal step. The pre-processing step of concept extraction takes considerable time. The label transfer step takes almost equal amount of time compared to JEC.

4 Conclusions

In this paper, we propose a novel knowledge-based approach for image annotation that exploits the semantic label concepts, derived based on the collective knowledge embedded in label co-occurrence based consistency network. An important future work to pursue is to build a *hierarchy of Concepts* and utilize them to learn useful insights for the tasks of image annotation and retrieval.

References

1. Makadia, A., Pavlovic, V., Kumar, S.: Baselines for image annotation. *International Journal of Computer Vision* **90**(1) (2010) 88–105
2. Verma, Y., Jawahar, C.V.: Image annotation using metric learning in semantic neighbourhoods. *ECCV* (2012) 836–849
3. Weston, J., Bengio, S., Usunier, N.: Wsabie : Scaling up to large vocabulary image annotation. *IJCAI* (2011)
4. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. *ICCV* (2009) 309–316
5. Jin, Y., Khan, L., Wang, L., Awad, M.: Image annotations by combining multiple evidence and wordnet. *ACM Multimedia* (2005)
6. Wang, M., Zhou, X., Xu, H.: Web image annotation based on automatically obtained noisy training set. *APWeb* (2008) 637–648
7. Opsahl, T., Agneessens, F., Skvoretz, J.: Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks* **32**(3) (2010) 245–251