Visual Recognition by Exploiting Latent Social Links in Image Collections

Li-Jia Li*[1], Xiangnan Kong*[2], Philip S. Yu[2]

1 Yahoo! Research, USA
2 Department of Computer Science, University of Illinois at Chicago
lijiali@yahoo-inc.com, {xkong4, psyu}@cs.uic.edu

Abstract. Social network study has become an important topic in many research fields. Early works on social network analysis focus on real world social interactions in either human society or animal world. With the explosion of Internet data, social network researchers start to pay more attention to the tremendous amount of online social network data. There are ample space for exploring social network research on large-scale online visual content. In this paper, we focus on studying multi-label collective classification problem and develop a model that can harness the mutually beneficial information among the visual appearance, related semantic content and the social network structure simultaneously. Our algorithm is then tested on CelebrityNet, a social network constructed by inferring implicit relationship of people based on online multimedia content. We apply our model to a few important multimedia applications such as image annotation and community classification. We demonstrate that our algorithm significantly outperforms traditional methods on community classification and image annotation.

1 Introduction

Visual recognition research by using image content has achieved promising progress in recent years. There emerges robust object detectors [1,2], large scale image classification methods [3,4], efficient image retrieval algorithms [5,6] and advanced image annotation algorithms [7,8], most of which are developed by using image content alone. At the same time, social network analysis has been playing an important role in many fields such as recommendation [9] and search [10]. In recent years, it has attracted significant amount of interests in the multimedia community [11,12,13]. Not much work has been done on exploiting the rich structural information in social networks, which cannot be captured by the traditional visual models, for the multimedia research. Such mutually beneficial information is very common in the large online multimedia data. For example, in Fig. 1, we can observe the consistency between the photos (visual content) and the tags (semantic information). In addition, people who are socially connected share similar photos and tags. Furthermore, photos and tags belonging to the same social group (indicated by the circle) are very characteristic. An interesting question is what is the benefit of introducing social network to high level visual recognition tasks. In this paper, we advocate that by modeling the mutual benefit of the visual content,

*indicates equal contributions.

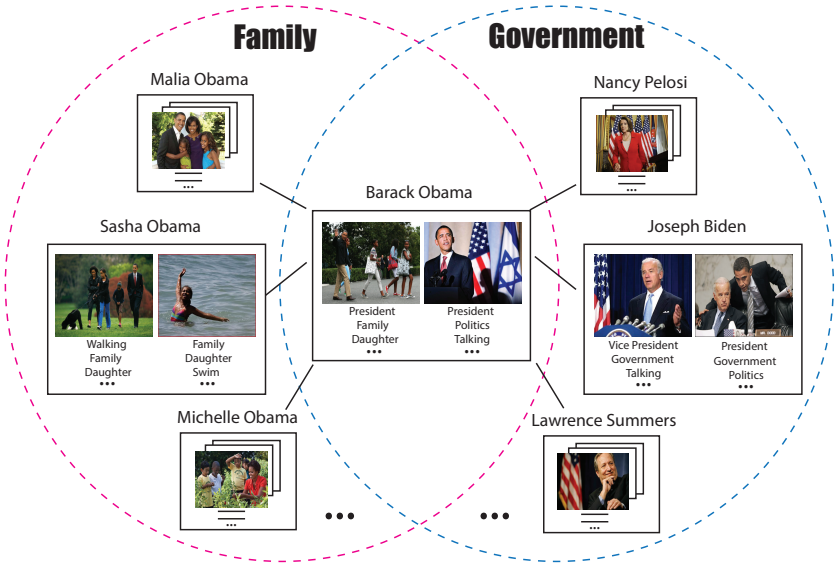


Fig. 1. Example to show the mutually beneficial information among photos, related tags and social structure. Each person has a set of photos and their related photo tags. Circles correspond to communities. Links between people indicate that they are connected.

semantic content and relationship structure in a social network, better visual recognition algorithm can be developed. Inspired by the ICA algorithm [14], we develop a collective classification algorithm to learn the mutually beneficial information modeled in this joint model. Fundamentally different than pioneering social network algorithms [11,12,13], our approach serves as generic annotation or classification algorithms which are not limited to face recognition or friendship prediction. We apply our algorithm to CelebrityNet, an implicit social network, as the source of visual, semantic and social structural information. It is constructed from the co-occurrence statistics of celebrities who appear in millions of professionally produced news images. It encodes visual, semantic and social structural information, making it a valuable resource for developing structural learning algorithms which jointly models such information. Experimental results demonstrate the effectiveness of our model built upon social network structure encoding multi-modal information resources for applications such as classification and image annotation by providing informative tags for unseen people or images. Specifically, we make the following contributions in this paper:

- Develop a principle model for modeling the mutually beneficial information among visual, semantic and implicit social structure from large scale online image dataset.
- Derive a collective classification algorithm for learning this model.
- Demonstrate significant result improvement by incorporating social structure in visual recognition.

It is worth noting that our model is a generic one. Although we demonstrate its effectiveness by using CelebrityNet in this paper, our algorithm is not limited to the

celebrity social network. It can be directly applied to other information sources, such as the users of a general social network or the objects of an online photo sharing website.

2 Related Works

Social network analysis has achieved substantial progress recently with the emergence of large scale online structural data [9,10]. While much of the research has been done on textual documents or hyper links, social network analysis on visual content has also made promising progress these years [15,11,12,13]. With the emergence of photo sharing websites such as Flickr! and Facebook, pioneer research [11,12] has been conducted to tackle visual recognition problems by incorporating social network structure. For example, [12] models the types of relationships based on face features such as face size ratio, age difference and gender distribution. [11] leverages the social network structure to improve face recognition in a collection of face photos by using a MRF model. Both models focus on face recognition and analysis, which are not directly applicable to generic recognition tasks. Little has been done to construct an implicit social network from visual data, uncover the structure embedded and applying it for generic multimedia tasks e.g. image annotation and classification. In this paper, we propose a model to harness the mutual information embedded in the social network encoding implicit relationship inferred from co-occurrence of people in photos.

At the mean time, visual recognition algorithms [1,2,3,4,7,8] have shown effectiveness in recognizing objects and classifying images. Most of them are using only the visual content of images. Interesting research such as the multi-label classification approaches [16,17] further explore the correspondence among the tags related to the images. Sophisticated models are developed to model the relationship between the visual content and the semantic meaning of the images. Our algorithm, on the other hand, takes the visual content, related semantic information and social network structure into account. We aim to emphasize the impact of social network structure in visual tasks as a new knowledge resource for multimedia tasks.

3 Implicit Social Network

Before we describe our model, we first introduce the social network constructed from an online image dataset and explain the motivation of developing our model. Intuitively, photos encode implicit relationship: People who are related to each other usually appear together and are photographed in some occasions; The more photos they appear in together, the stronger the relationship might be. Inspired by this observation, we construct a social network called CelebrityNet from large scale online celebrity image dataset collected from a professional photograph website [1] named Getty [18]. In this social network, a network link is placed between two persons if they appear in images together. The strength of the link is simply the frequency of their occurrence in the dataset.

[1] Person names are manually labeled in this image dataset. Specifically, we use 2 million images to construct our implicit social network.

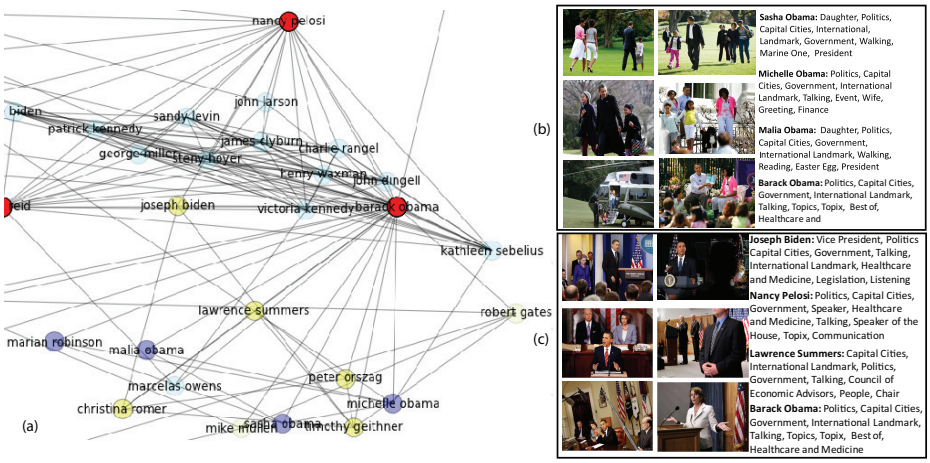


Fig. 2. Sample overlapped communities, related photos and popular tags. (a) Overlapped communities of 'Barack Obama'. Nodes represent people in the social network. Lines denote connection between people. Nodes highlighted in the same color refers to people belonging to the same community. Persons assigned to multiple communities are highlighted in red. (b) Example images and frequent tags of each person in the 'Obama Family' community. (c) Example images and popular tags of each person in a 'Obama Government' community.

Human beings, as social creatures, naturally form communities due to similar profession, location, and hobbies etc., which can be reflected by photos they take together. Let's take a deep dive in the constructed social network and uncover such phenomena. Our assumption is that the group of people belonging to a community often appears together much more often than people who are not part of the group. At the same time, one person could belong to multiple communities. For example, a person with computer vision research as profession could have the hobby of cooking and he/she then simultaneously belongs to both the computer vision researcher community and the cooking community. Therefore, we adopt the Clique Percolation Method (CPM) proposed by Palla et al. [19] to discover the overlapping communities.

In Fig. 2, we show example of overlapped communities related to 'Barack Obama' with the images and the most frequent tags for the persons in the community. From Fig. 2 , our first observation is the consistency between visual appearance and the tags. Traditional image annotation algorithms [1,2,3] are developed based on this observation. By modeling the correspondence between the visual and semantic content, these algorithms are capable of automatically predicting related tags for unknown test images. In addition, people who are linked to each other in the social network usually share visual and semantic content. Furthermore, we observe that images and tags are very characteristic within each community. This applies not only to isolated communities but also overlapped communities. Social network encodes informative structure for learning the visual and textual data. Inspired by this observation, we propose a model jointly models the visual, textual and social information in Section 4.

4 Models

In this paper, we formulate the image annotation and community classification tasks as multi-label classification problems. For each unknown test image, our algorithm needs to provide a list of tags related to it. Similarly, each person in the social network will be assigned to one or multiple communities. We first introduce some notations that will be used throughout. Suppose we have a multi-label dataset $\mathcal{D}(\mathcal{X}, \mathcal{Y}) = \{(\boldsymbol{x}_i, \boldsymbol{Y}_i)\}_{i=1}^n$ and a network $G(\mathcal{V}, \mathcal{E})$ among the samples of the dataset. Here $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^n$, and $\boldsymbol{x}_i \in \mathbb{R}^d$ denotes the feature vector of sample $\boldsymbol{x}_i$ in the d-dimensional input space. $\mathcal{Y} = \{\boldsymbol{Y}_i\}_{i=1}^n$, where $\boldsymbol{Y}_i = \left(Y_i^1, \cdots, Y_i^q\right)^\top \in \{0,1\}^q$ denotes the multiple labels assigned to sample $\boldsymbol{x}_i$. Let $\mathcal{C} = \{\ell_1, \cdots, \ell_q\}$ denote the set of q possible label concepts. In the network G, $\mathcal{V} = \{v_1, \cdots, v_n\}$ is a set of nodes, which corresponds to the samples in $\mathcal{D}$. $\mathcal{E}$ is the set of links/edges in $\mathcal{V} \times \mathcal{V}$. Assume that we have a training set $\mathcal{X}_\mathcal{L} \subset \mathcal{X}$ where the values $\mathcal{Y}_\mathcal{L}$ are known. Here $\mathcal{L}$ denotes the index set for training data, *i.e.*, $\mathcal{Y}_\mathcal{L} = \{\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{x}_i \in \mathcal{X}_\mathcal{L}\}$. $\boldsymbol{y}_i = \left(y_i^1, \cdots, y_i^q\right)^\top \in \{0,1\}^q$ is a binary vector representing the observed label set assigned to sample $\boldsymbol{x}_i$. $y_i^k = 1$ if the k-th label is in $\boldsymbol{x}_i$'s label set.

Multi-label collective classification corresponds to the task of predicting the values of all $\boldsymbol{Y}_i \in \mathcal{Y}_\mathcal{U}$ for the testing set collectively ($\mathcal{X}_\mathcal{U} = \mathcal{X} - \mathcal{X}_\mathcal{L}$), where the inference problem is to estimate $\Pr(\mathcal{Y}_\mathcal{U} | \mathcal{X}, \mathcal{Y}_\mathcal{L})$. Conventional supervised classification approaches usually has *i.i.d.* assumptions, *i.e.*, the inference for each sample is independent from other samples, *i.e.*, $\Pr(\mathcal{Y}_\mathcal{U} | \mathcal{X}, \mathcal{Y}_\mathcal{L}) \propto \prod_{i \in \mathcal{U}} \Pr(\boldsymbol{Y}_i | \boldsymbol{x}_i)$. Moreover, in multi-label classification, the simplest solution (*i.e.*, one-vs-all) assumes that the inference of each label is also independent from other labels for an sample, *i.e.*, $\Pr(\boldsymbol{Y}_i | \boldsymbol{x}_i) = \prod_{k=1}^q \Pr(Y_i^k | \boldsymbol{x}_i)$. However, in many real-world classification tasks, there are complex dependencies not only among different samples but also among different labels.

In order to solve the multi-label collective classification problem more effectively, we explicitly consider three types of relationships. We adopt the multi-kernel learning framework (MKL) [20] and build one kernel on each type of relationship. SVMs have been widely used for classification problems in recent years. Different than traditional SVM, MKL incorporates multiple kernels and can learn a convex combination of these kernels (*i.e.*, the kernel weights) simultaneously $\mathbf{K} = \sum_i \beta_i \mathbf{K}_i$. Specially, we build three different kernels that can capture three different types of relationship in the data.

4.1 Content Relationship

The first type of relationships we consider is about the visual content features of the samples. Conventional image annotation approaches focus on using the image content features to build inference models. In order to capture the content/visual information of different samples, we build the *content kernel* based upon the input visual feature vector of different samples. $K_{content}(i,j) = \phi(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Here, any conventional kernel function can be used for $\phi(\cdot,\cdot)$. Intuitively, the content kernel denotes the relationship that if two images share similar visual features, they are more likely to have similar labels.

Input:
$\mathcal{G}$: a network, $\mathcal{X}$: attribute vectors for all instances.
$\mathcal{Y}_{\mathcal{L}}$: label sets for the training instances, A: a base learner for multi-kernel learning model, $T_{\max}$: maximum # of iteration (default=10)
Training:
- Learn the MKL model f:
  1. Construct q extended training sets $\forall 1 \leq k \leq q, \mathcal{D}_k = \left\{(\mathbf{x}_i^k, y_i^k)\right\}$ by converting each instance $\boldsymbol{x}_i$ to $\mathbf{x}_i^k$ as follows:
     $\boldsymbol{x}_i^k = (\boldsymbol{x}_i, \text{LabelSetFeature}(\ell_k, \boldsymbol{Y}_i), \text{NetworkFeature}(i, \mathcal{Y}_{\mathcal{L}}))$
  2. Computer the corresponding kernels for each label: Φ, $\Phi_{Labelset}$, and $\Phi_{network}$
  3. Calculate kernel weights and train MKL models on each label. Let $f_k = A(\mathcal{D}_k)$ be the MKL model trained on $\mathcal{D}_k$.
Bootstrap:
- Estimate the label sets, for $i \in \mathcal{U}$: produce an estimated values $\hat{\boldsymbol{Y}}_i$ for $\mathbf{Y}_i$ as follows: $\hat{Y}_i = f\left((\boldsymbol{x}_i, \mathbf{0})\right)$ using attributes only.
Iterative Inference:
- Repeat until convergence or #iteration$> T_{\max}$
  1. Construct the extended testing instance by converting each instance $\boldsymbol{x}_i$ to $\boldsymbol{x}_i^k$'s ($i \in \mathcal{U}$) as follows:
     $\boldsymbol{x}_i^k = \left(\boldsymbol{x}_i, \text{LabelsetFeature}(\ell_k, \hat{\boldsymbol{Y}}_i), \text{NetworkFeature}(i, \mathcal{Y}_{\mathcal{L}} \cup \{\hat{\boldsymbol{Y}}_i | i \in \mathcal{U}\})\right)$
  2. Update the estimated value $\hat{\boldsymbol{Y}}_i$ for $\boldsymbol{Y}_i$ on each testing instance ($i \in \mathcal{U}$) as follows: $\forall 1 \leq k \leq q$, $\hat{Y}_i^k = f_k(\boldsymbol{x}_i^k)$.
Output:
$\hat{\mathcal{Y}}_{\mathcal{U}} = \left(\hat{\boldsymbol{Y}}_1, \cdots, \hat{\boldsymbol{Y}}_{n_u}\right)$: the label sets of testing instances ($i \in \mathcal{U}$).

Fig. 3. The MKML algorithm

4.2 Label Set Relationship

The second type of relationships we consider is the label correlations within the label set of each sample. Different labels are inter-dependent in multi-label classification, thus should be predicted collectively. For example, in image annotation tasks, an image is more likely to have the tag 'sports' if we know the image has already been assigned with the tag 'NBA' or 'basketball'. The image is less likely to be annotated as 'sports', if we already know the image contains the label 'academy awards'.

Conventional multi-label classification approaches focus on exploiting such label correlations to improve the classification performances, which model $\Pr(Y_i^k | \boldsymbol{x}_i, \boldsymbol{Y}_i^{\{-k\}})$. $\boldsymbol{Y}_i^{\{-k\}}$ represents the vector of all the variables in the set $\{Y_i^p : p \neq k\}$. Hence, we have $\Pr(\boldsymbol{Y}_i | \boldsymbol{x}_i) = \prod_{k=1}^{q} \Pr(Y_i^k | \boldsymbol{x}_i, \boldsymbol{Y}_i^{\{-k\}})$. Based upon the above observation, we build the *label set kernel* encoding the correlations among different labels. $K_{labelset}(Y_i^k, Y_j^k) = \phi\left(\boldsymbol{Y}_i^{\{-k\}}, \boldsymbol{Y}_j^{\{-k\}}\right)$. Intuitively, the label set kernel denotes the relationship that if two images share similar label sets, they are more likely to have similar values in any label variable.

4.3 Network Relationship

The third type of relationships we consider is the correlations among label sets of the related samples that are inter-connected in the network. The label sets of related samples are usually inter-dependent in a network. For example, in our CelebrityNet network, the probability of an image having the label 'politics' should be higher if we already know the image contains the same people appearing in some other images with a label set of {'government', 'politics'}.

Conventional collective classification approaches focus on exploiting this type of dependencies to improve the classification performances, which models $\Pr(Y_i^k | \boldsymbol{x}_i, \boldsymbol{Y}_{j \in \mathcal{N}(i)})$.

Here $\boldsymbol{Y}_{j\in\mathcal{N}(i)}$ denotes the set containing all vectors $\boldsymbol{Y}_j$ ($\forall j \in \mathcal{N}(i)$), and $\mathcal{N}(i)$ denotes the index set of related samples to the i-th sample, *i.e.*, the samples directly linked to the i-th sample. Hence, we will have $\Pr(\boldsymbol{Y}_{\mathcal{U}}^k|\boldsymbol{X}) = \prod_{i\in\mathcal{U}} \Pr(Y_i^k|\boldsymbol{x}_i, \boldsymbol{Y}_{j\in\mathcal{N}(i)})$. Based upon the above observation, we build the *network kernel* encoding the correlations among related samples that are connected in the network. $K_{network}(Y_i^k, Y_j^k) = \phi\left(\boldsymbol{Y}_{l\in\mathcal{N}(i)}, \boldsymbol{Y}_{l\in\mathcal{N}(j)}\right)$. Intuitively, the network kernel denotes the relationship that if the neighbors of the two images share similar label sets, these two images are more likely to have similar label sets.

The general idea is as follows: We build one kernel on each type of the relations mentioned above, and then use MKL method to learn the weights of the multiple kernels (i.e., the importance of different kernels). We model the joint probability based upon the Markov property: if sample $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are not directly connected in network G, the label set $\boldsymbol{Y}_i$ is conditional independent from $\boldsymbol{Y}_j$ given the label sets of all $\boldsymbol{x}_i$'s neighbors. The local conditional probability on label k can be modeled by a MKL learner with aforementioned kernels. The computation of these kernels depends on the predicted $\boldsymbol{Y}_j$ ($j \in \mathcal{N}(i)$) and the predicted $\boldsymbol{Y}_i^{\{-k\}}$. Then, the joint probability can be approximated based on these local conditional probabilities by treating different labels as independent and the samples as *i.i.d.*. To simply demonstrate the effectiveness of our approach, we use linear kernels for all relations here.

Motivated by the ICA framework [14,21], we proposed the following inference procedure of our MKML method as shown in Figure 3. (1) At the beginning of the inference, the label sets of all the unlabeled samples are unknown. The *bootstrap* step is used to assign an initial label set for each sample using the content feature of each sample. In our current implementation, we simply initialize the label set features and the network features for unlabeled samples with all zero vectors. Other strategies can also be used for bootstrapping, e.g, training SVM (single kernel) on training data using content feature only, and then we use these models to assign the initial label sets of unlabeled samples. (2) In the *iterative inference* step, we iteratively update the label set features/kernels and network features/kernels based upon the predictions of MKL models and update the prediction of MKL models using the newly updated kernels. The iterative process stops when the predictions of all MKL models are stabilized or a maximum number of iteration has been reached.

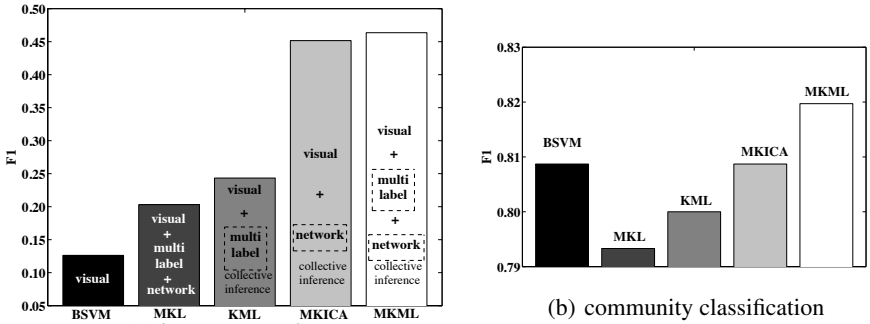
5 Experiment

5.1 Compared Methods

In this subsection, we compared a set of methods exploring different information resources:

• BSVM (binary SVM): This baseline method uses binary decomposition to train one classifier on each label separately, which is similar to [22]. BSVM assumes all the labels and all instances are independent. It is based on visual content alone.

• MKL (Multi-kernel learning): We directly apply multiple kernel learning algorithm on the joint information of the visual, semantic and social network without iterative inference steps.



(a) image annotation (b) community classification

Fig. 4. Overall performances of the compared methods.

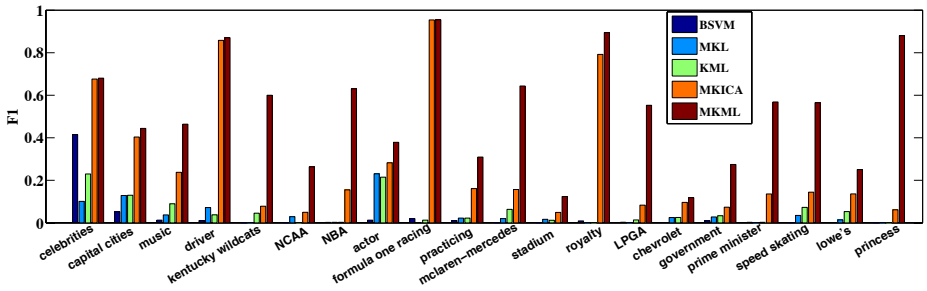


Fig. 5. F1 scores on example labels in image annotation task.

•KML (visual kernel + multi-label kernel): This baseline method trains one multi-kernel learner on each label, using two different kernels visual feature kernel and multi-label kernel. KML not only models the correspondence between the visual content and the tags, but also models the correlation among the tags.

• MKICA (visual kernel + network kernel): In this baseline method, the multi-label dataset is first divided into multiple single-label datasets by one-vs-all binary decomposition. For each binary classification task, we use a multi-kernel version of ICA [14], as the base classification method. MKICA combines the social structure with visual modeling of the tags. However, it ignores the relationship among the tags.

•MKML (Multi-kernel Multi-label Collective classification): Our proposed method for multi-label collective classification based upon multi-kernel learning, which jointly models the visual, semantic and social network information.

For a fair comparison, we use LibLinear [23] as the base classifier for BSVM and LibLinear MKL as the base learner for all the remaining methods. The maximum number of iterations in the methods KML, MKICA, and MKML are all set as 10 based on observation from the validation experiment.

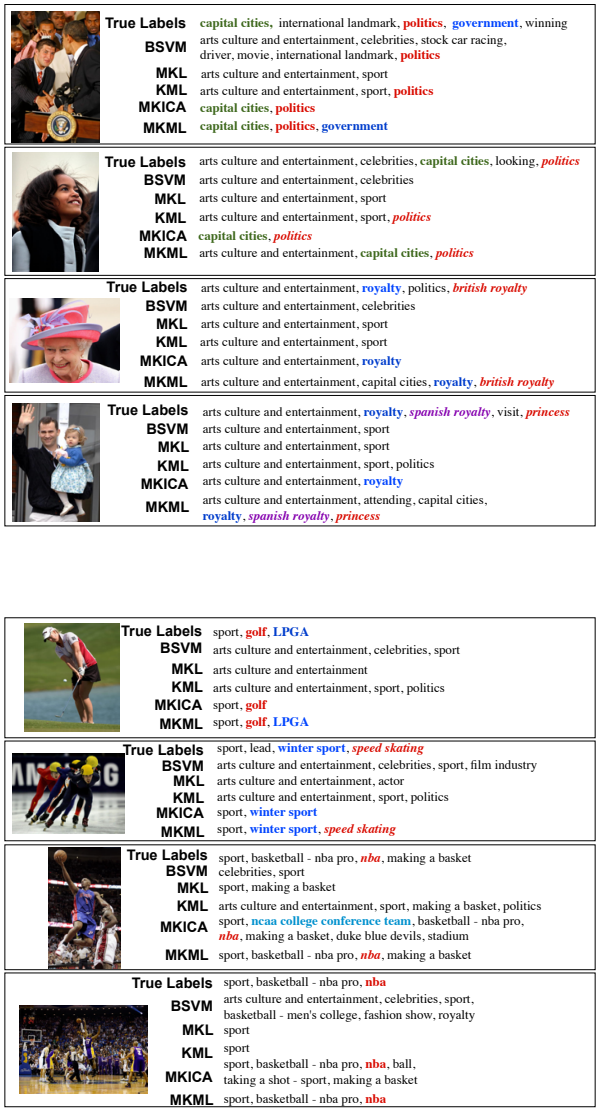


Fig. 6. Annotation Examples of different algorithms. BSVM, MKL, KML, MKICA, MKML represent binary SVM, traditional multi-kernel learning method, method built upon visual kernel + network kernel, Multi-kernel Multi-label Collective classification method respectively. Tags exclusively recognized by our methods are highlighted in color.

5.2 Evaluation Metrics

We use some evaluation criteria in [24,25,26] to verify the image annotation performance. Suppose a multi-label dataset $\mathcal{D}_U$ contains n instances $(\boldsymbol{x}_i, \boldsymbol{Y}_i)$, where $\boldsymbol{Y}_i \in \{0,1\}^q$ $(i = 1, \cdots, n)$. Denote $h(\boldsymbol{x}_i)$ as the predicted label set for $\boldsymbol{x}_i$ by a multi-label classifier h, we have

• *F1*: is the harmonic mean of precision and recall.

$$\mathrm{F1}(h, \mathcal{D}_U) = \frac{2 \times \sum_{i=1}^{n} \|h(\boldsymbol{x}_i) \cap \boldsymbol{Y}_i\|_1}{\sum_{i=1}^{n} \|h(\boldsymbol{x}_i)\|_1 + \sum_{i=1}^{n} \|\boldsymbol{Y}_i\|_1}$$

The larger the value, the better the performance.

All experiments are conducted on a machine with Intel Xeon$^{\mathrm{TM}}$Quad-Core CPUs of 2.26 GHz and 24 GB RAM. We tested the performances on the following tasks:
1) image annotation task: we have 102,565 images with 159 frequent tags, where each image can be annotated with a subset of these tags. On each image we extracted 5000 dimensional visual features in bag-of-words representation. We then randomly sample two thirds of the images as the training set, and use the remaining images as the test set.
2) community classification: we have 554 people in the dataset, where each person can be classified into a subset of 80 candidate communities. We randomly sample 436 people into the training set, and use the remaining 118 people as the test set. For each person, we use the aggregated visual features of all his/her photos. Two persons are linked together if they appeared in at least one photo.

5.3 Results

As mentioned earlier in the paper, visual content, semantic information and the social network structure are mutually beneficial to each other. Below, we demonstrate results of two visual recognition tasks to show the advantage of jointly modeling these three information sources. Specifically, image annotation task illustrates the potential of our approach for predicting semantic information based on visual content and the social network structure. Community classification of unknown person based on his/her set of photos and related tags demonstrates the possibility of using visual and semantic tags for social network structure prediction.

In Fig. 4(a), we make the following observations: 1. The visual content based approach BSVM achieves reasonably good performance in image annotation [2], indicating strong correlation between the visual content and the tags. 2. Learning the correlation among tags is helpful, reflected by the improvement of KML over BSVM. This improvement is understandable: a photo with tag 'NBA' usually has 'basketball' in the tag list as well. 3. Methods MKICA and MKML significantly outperform the other methods indicating that incorporating the social network structure is especially useful. From the analysis of social network in Section 3, we learn that images and tags belonging to the same person and same community are very characteristic. Therefore, modeling the social network structure naturally improve the tag prediction performance of unknown images. 4. The significant improvement of MKML over the traditional MKL shows the

[2] Random approach achieves only 0.03 by using the F1 measure.

power of iterative prediction and error correction in our proposed method. 5. Finally, jointly modeling the visual, semantic and social structure (MKML) provides additional improvement over combining visual and semantic information (MKL), demonstrating the effectiveness of social network structure.

In Fig. 4(b), we show the community classification results of different algorithms. In this experiment, tag correlation (KML) is not as useful as it is in the image annotation task. This is interpretable: as long as we know the person is related to the tag 'NBA', we can already do a good community classification without knowing other tags. On the other hand, if we know whom the unknown person is connected to, it is fairly easy to predict his/her community. This leads to the good performance of social network based algorithms.

To provide more details of the annotation result, we show the F1 scores of example labels in Fig. 5. While we observe similar pattern as in Fig. 4(a) with clear advantage of the social network based algorithms over the other methods, the social network based algorithms usually perform much better on specific labels with social meaning such as 'kentucky wildcats' and 'royalty'. Such social meaning can not be inferred from the visual content. The observation aligns well with our motivation of incorporating social network structure as a source of complimentary information for high level visual recognition tasks.

Finally, we show example results of image annotation in Fig. 6. Visual only method provides conservative prediction of common tags correlated to the visual content. Incorporating social network upon the visual and semantic modeling enables the algorithms to be more accurate in image annotation. MKML further enrich the tag annotation accurately by exploring the tag correlation upon jointly modeling the three sources of information. For example, in the 4th picture, the tags 'princess' and 'Spanish royalty' can only be inferenced correctly by combining information from Social netowrks and correlations with other tags (such as 'royalty').

6 Conclusion

In this paper, we propose a model to jointly model the visual content, semantic information and relationship structure for a few multimedia tasks. Our algorithm has been tested on a social network constructed from large scale images. We demonstrate significant improvement on community classification and image annotation of celebrity images over related algorithms. Our algorithm is a generic algorithm for modeling mutual information of content and relational data. In the future, we would like to explore its potential on generic online user photos such as those available on Flickr! for automatically prediction of missing tags. Another interesting direction is to develop advanced algorithms upon the proposed one for high level visual recognition tasks such as large scale object detection by combining detailed visual content information and objects' relational structure.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005) 886 1, 3, 4

2. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object Detection with Discriminatively Trained Part Based Models. JAIR **29** (2007) 1, 3, 4
3. Lin, Y., Lv, F., Zhu, S., Yang, M., Cour, T., Yu, K., Cao, L., Huang, T.: Large-scale image classification: fast feature extraction and svm training. In: CVPR. (2011) 1689–1696 1, 3, 4
4. Deng, J., Berg, A., Li, K., Fei-Fei, L.: What does classifying more than 10,000 image categories tell us? ECCV (2010) 71–84 1, 3
5. Wu, Z., Ke, Q., Isard, M., Sun, J.: Bundling features for large scale partial-duplicate web image search. In: CVPR. (2009) 25–32 1
6. Jégou, H., Douze, M., Schmid, C.: Improving bag-of-features for large scale image search. IJCV **87** (2010) 316–336 1
7. Cao, L., Yu, J., Luo, J., Huang, T.: Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression. In: ACM MM. (2009) 1, 3
8. Weston, J., Bengio, S., Usunier, N.: Large scale image annotation: Learning to rank with joint word-image embeddings. Machine learning **81** (2010) 21–35 1, 3
9. Konstas, I., Stathopoulos, V., Jose, J.: On social networks and collaborative recommendation. In: SIGIR, ACM (2009) 195–202 1, 3
10. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems **30** (1998) 107–117 1, 3
11. Stone, Z., Zickler, T., Darrell, T.: Toward large-scale face recognition using social network context. Proceedings of the IEEE **98** (2010) 1408–1415 1, 2, 3
12. Wang, G., Gallagher, A., Luo, J., Forsyth, D.: Seeing people in social context: Recognizing people and social relationships. In: ECCV. (2010) 1, 2, 3
13. Zhuang, J., Mei, T., Hoi, S., Hua, X., Li, S.: Modeling social strength in social media community via kernel-based learning. In: ACM MM. (2011) 1, 2, 3
14. Lu, Q., Getoor, L.: Link-based classification. In: ICML. (2003) 2, 7, 8
15. Ding, L., Yilmaz, A.: Learning relations among movie characters: A social network perspective. ECCV (2010) 410–423 3
16. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. In: ECML/PKDD. (2009) 3
17. Liu, X., Shi, Z., Li, Z., Wang, X., Shi, Z.: Sorted label classifier chains for learning images with multi-label. In: ACM MM. (2010) 3
18. http://www.gettyimages.com/. 3
19. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435** (2005) 814–818 4
20. Vishwanathan, S.V.N., Sun, Z., Theera-Ampornpunt, N.: Multiple kernel learning and the smo algorithm. In: NIPS. (2010) 5
21. McDowell, L.K., Gupta, K.M., Aha, D.W.: Cautious inference in collective classification. In: AAAI, Vancouver, Canada (2007) 596–601 7
22. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition **37** (2004) 1757–1771 7
23. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: a library for large linear classification. (2008) 8
24. Ghamrawi, N., McCallum, A.: Collective multi-label classification. In: CIKM, Bremen, Germany (2005) 195–200 10
25. Kang, F., Jin, R., Sukthankar, R.: Correlated label propagation with application to multi-label learning. In: CVPR, New York, NY (2006) 1719–1726 10
26. Liu, Y., Jin, R., Yang, L.: Semi-supervised multi-label learning by constrained non-negative matrix factorization. In: AAAI, Boston, MA (2006) 421–426 10