2012 IEEE 12th International Conference on Data Mining

A Classification Based Framework For Concept Summarization

Dhruv Mahajan*, Sundararajan Sellamanickam†, Subhajit Sanyal‡ and Amit Madaan§
*Microsoft Research India, email: dhrumaha@microsoft.com
†Microsoft Research India, email: ssrajan@microsoft.com
‡American Express Bangalore, email: subhajit.sanyal@gmail.com
§TheFind, Mountain View, amadaan@thefind.com

***Abstract*—In this paper we propose a novel classification based framework for finding a small number of images that summarize a given concept. Our method exploits metadata information available with the images to get category information using Latent Dirichlet Allocation. Using this category information for each image, we solve the underlying classification problem by building a sparse classifier model for each concept. We demonstrate that the images that specify the sparse model form a good summary. In particular, our summary satisfies important properties such as likelihood, diversity and balance in both visual and semantic sense. Furthermore, the framework allows users to specify desired distributions over categories to create personalized summaries. Experimental results on seven broad query types show that the proposed method performs better than state-of-the-art methods.**

***Keywords*-concept summarization; classification; metadata;**

I. Introduction

The problem of summarization arises in the context of various applications for different media like images, photo collections and videos. Examples of summarization include (a) presenting a *slide show* on an event like Oil Spill from a collection of images and text descriptions collected from various sources, and (b) displaying a diverse yet relevant collection of images for an image search query. With the amount of data growing enormously, this problem has received considerable attention in the computer vision, multimedia and web communities.

This problem is challenging because an effective summarization should have some important properties [1], [2], [3], [4] from both the semantic and visual viewpoints. The three properties commonly used and considered in this work are:
Diversity - No two images in the summary should be similar to each other visually or semantically.
Likelihood - An image in the summary should be similar to many other images in the dataset visually or semantically. Note that commonly occurring visual and semantic aspects are generally more relevant.
Balance - The various visual and semantic aspects should be present in a balanced way to avoid any misunderstanding of the concept summary.

Additionally, there could be user specified constraints like maximum number of images, preference to particular semantic aspects, time scale, etc., that a summary should satisfy. Consider a concept like Oil Spill (Fig. 1). It has different aspects such as environmental impact, protests, etc. Showing images belonging to various semantic aspects of a concept is a key requirement in concept summarization.

There exists a large collection of work on summarization in the literature [1], [3], [4], [5], [6], [7], [8] varying from application viewpoint, information and media used, etc. However, all of them have issues with either maintaining a balance between semantic and visual aspects or trading-off diversity with likelihood. We discuss these issues in the related work section.

In this work, we focus on the *problem of summarizing images for a given concept* when additional information is available in the form of metadata. The motivation behind our approach is based on two observations: (1) text description of an image often gives important information about semantic aspects (topics), and (2) image features are often correlated with semantic topics. The goal is to discover various aspects of the concept and present a representative set of images covering these aspects. To discover the semantic aspects, we build a Latent Dirichlet Allocation (LDA) model [9] using text descriptions of all the images. We then build a sparse classifier model (specified by a subset of images forming the summary) using the discovered categories as classes, to correlate the discovered categories with the visual features. In particular, the sparse model designed using likelihood maximization principle ensures that chosen images maintain a good trade-off between visual properties while covering all the semantic aspects effectively.

Our Contributions

1) We propose a novel, classification based framework for solving the problem of concept summarization. Although our approach is classification based, it is *still unsupervised*. This is because we *automatically* get the category information of the images from the topics discovered by LDA.
2) Our framework allows us to specify distributions over the categories that the summary should satisfy, in a simple fashion. These distributions give the flexibility to customize summaries for different users. To the best of our knowledge, this has never been attempted before.
3) Since qualitative evaluation of summarization can be subjective, we additionally propose a set of quantitative metrics to compare the performance of different summarization methods on various visual and semantic aspects. These

1550-4786/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDM.2012.114

Figure 1. Concept: Oil Spill. Different semantic aspects: A - Env. (Sea,Beach), B - Env. (Birds and Fish), C - Corporate, D - Protests, E - Politics.

metrics are also useful to compare the methods on a large collection of datasets as manual evaluation is expensive.

II. Our Approach

Given a concept, we first get a relevant collection of images with text descriptions (metadata). Such collections can be easily obtained from different sources such as *flickr*, or, feeds from the sites like http://news.yahoo.com. Using these inputs, we build a sparse classifier model specified by a subset of images with class information obtained by building an LDA model on the text.

A. Classification Framework

In this section, we explain how the subset of images (summary) is obtained from building a sparse classifier model. Let $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$ and $\mathcal{T} = \{T_1, T_2, \ldots, T_N\}$ denote a collection of images with respective text descriptions. Thus, a concept is represented by the tuple $(\mathcal{I}, \mathcal{T})$. Let us assume that $\mathcal{X} = \{\mathbf{x}_i : i = 1, \ldots, N\}$ is a feature representation of $\mathcal{I}$. Let us denote the topic distributions of the collection $\mathcal{I}$ obtained from the metadata $\mathcal{T}$ using LDA as: $\mathcal{Q} = \{\mathbf{q}_i : i = 1, \ldots, N\}$, where $\mathbf{q}_i = [q_{i,1}, \ldots, q_{i,M}]$ is the topic distribution of the i^{th} image and M denotes the number of LDA topics. Usually, each image belongs to very few topics. Therefore, $\mathbf{q}_i$ is sparse. Our idea is to treat each topic as a class and $\mathbf{q}_i$ as the class probability distribution over a set of classes $C = \{c_1, c_2, \ldots, c_M\}$, i.e., $P(c_j|\mathbf{x}_i) = q_{i,j}, \forall i = 1, \ldots, N, j = 1, \ldots, M$. Thus, the problem is: *given the image-class distribution pairs for all the examples, find a sparse set of images $\mathcal{I}_{\mathbf{S}} \subset \mathcal{I}$ that summarizes the concept for some $\mathbf{S} \subset \{1, \ldots, N\}$ with a user specified value L, where $L = |\mathbf{S}| \ll N$.* As shown below, we map this problem to the problem of designing a sparse kernel classifier using likelihood maximization principle with $(\mathcal{X}, \mathcal{Q})$ as the input.

Several sparse kernel classifier design methodologies have been proposed in the literature [10]. In the binary setting, a sparse kernel classifier decision function can be defined as: $f(\mathbf{x}) = \sum_{i \in \mathbf{S}} a_i k(\mathbf{x}, \mathbf{x}_i)$, where each $k(\mathbf{x}, \mathbf{x}_i)$, a_i are referred as a basis function and its coefficient respectively. It is well known [10] that sparse classifier models achieve classification accuracy closer to that achievable by the full model (using all the examples instead of $\mathbf{S}$), when $\mathbf{S}$ is carefully chosen according to some suitable criterion. We conjecture that such a subset ($\mathbf{S}$) forms a good summary. For our purpose, we are interested in sparse probabilistic kernel models. This is because the class distribution information $\mathcal{Q}$ that we get from LDA can be naturally incorporated in the probabilistic framework. We use import vector machine (IVM) [10] as the sparse kernel classifier in this work.

Import Vector Machine: Let c_i denote the class label of the i^{th} image and $\mathbf{y}_i$ be a M-dimensional binary vector with only one non-zero element at c_i^{th} position. Then, the IVM optimization problem for a multi-class classification problem can be written as [10]:

$$\min_{\mathbf{a},\mathbf{S}} \; -\frac{1}{N}\sum_{i=1}^{N} y_{i,c_i} \log(P(c_i|\mathbf{x}_i)) + \frac{\lambda}{2}\sum_{m=1}^{M} \mathbf{a}_{:,m}^T \mathbf{K} \mathbf{a}_{:,m} \quad (1)$$

$$P(c_j|\mathbf{x}) = \frac{e^{f_j(\mathbf{x})}}{\sum_{m=1}^{M} e^{f_m(\mathbf{x})}}, \quad f_m(\mathbf{x}) = \sum_{i \in \mathbf{S}} a_{i,m} k(\mathbf{x}, \mathbf{x}_i) \quad (2)$$

Result: Subset $\mathbf{S}$ of indices of the summary images

begin
  $\mathbf{S} \longleftarrow \emptyset$
  for $k \leftarrow 1$ **to** L **do**
    $\bar{\mathbf{S}} \longleftarrow \{1, ..., N\} \backslash \mathbf{S}$
    for $i \in \bar{\mathbf{S}}$ **do**
      Construct $\mathbf{S}_i \longleftarrow \mathbf{S} \cup \{i\}$.
      Set $\mathbf{S}$ to $\mathbf{S}_i$ in Equation 3 and optimize over $\mathbf{a}$ using gradient descent algorithm
      $E_i \leftarrow$ optimized objective function value in Equation 3
      if *Distribution constraint is applicable* **then**
        $E_i \longleftarrow E_i + \eta \mathrm{KL}(P_t, \frac{1}{|\mathbf{S}_i|} \sum_{l \in \mathbf{S}_i} \mathbf{q}_l)$.
      end
    end
    $\mathbf{S} \longleftarrow \mathbf{S} \cup \{\arg\min_{i \in \bar{\mathbf{S}}} E_i\}$.
  end
  Do finer optimization of the coefficients $\mathbf{a}$ using gradient descent algorithm
end

Algorithm 1: A Greedy Algorithm for Subset Selection

where λ is a regularization constant and $(i,j)^{\text{th}}$ entry of the matrix $\mathbf{K}$ is given by $k(\mathbf{x}_i, \mathbf{x}_j)$. $\mathbf{a}_{:,m}$ denotes the coefficient vector of the m^{th} classifier, $f_m(\mathbf{x})$. The IVM model gives a natural estimate of probability that an example $\mathbf{x}_i$ belongs to class c_j (Eq. 2). In IVM, the examples in $\mathbf{S}$ are called the import vectors, and we use these import vectors as the concept summary $(\mathcal{I}_{\mathbf{S}}, \mathcal{T}_{\mathbf{S}})$.

In our general setting, we have the probability distribution $\mathbf{q}_i$ (instead of $\mathbf{y}_i$) where more than one class can have nonzero values. To handle this, we extend (Eq. 1) to

$$\min_{\mathbf{S},\mathbf{a}} \; -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} q_{i,j} \log(P(c_j|\mathbf{x}_i)) + \frac{\lambda}{2} \sum_{m=1}^{M} \mathbf{a}_{:,m}^T \mathbf{K} \mathbf{a}_{:,m}. \tag{3}$$

Note that the first term is equivalent to minimizing the Kullback-Leibler (KL) divergence between the target distribution $\mathbf{q}_i = [q_{i,1} q_{i,2} \cdots q_{i,M}]$ and the model distribution $\mathbf{p}(c|\mathbf{x}_i) = [p_{i,1} p_{i,2} \cdots p_{i,M}]$ induced by the IVM model. Thus, the classifier model is built to predict the topic distribution of an image, thereby connecting the (visual) image features with (semantic aspects) topics. The second term is a regularization term. Note that Eq. 3 is a combinatorial optimization problem in $\mathbf{S}$. Therefore, we use a greedy algorithm (Algorithm 1).

B. Choice of Import Vectors

We now explain why the import vectors satisfy the visual and semantic properties presented in Section I.

Firstly, in order to model the class likelihood function in Eq. 3, we need the decision functions $f_m(\mathbf{x})$ (Eq. 2)

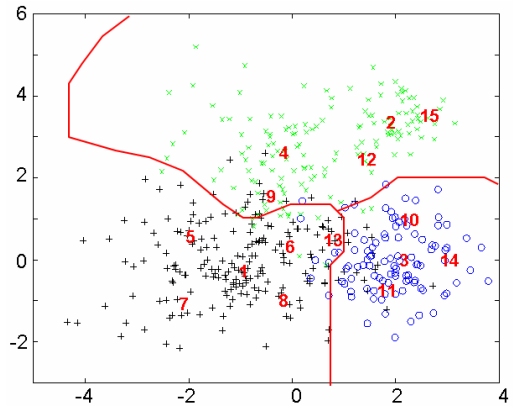

Figure 2. Import Vectors for 3-class Mixture of Gaussian Example. Please see the color image for better clarity.

belonging to all classes estimated well. This requires selection of images from all the classes representing different semantic aspects. Such a selection ensures semantic diversity. We illustrate this through an example. Fig. 2 shows a toy 4-component Gaussian mixture data belonging to three categories. Categories 1 (o - blue) and 3 (+ - black) consists of a single Gaussian, where as category 2 (x - green) consists of a mixture of two Gaussians. We show the top 15 import vectors numbered as per the selection order at their respective positions. Note that the first 3 import vectors come from three different classes, ensuring semantic diversity. Next, observe that the first import vector comes nearly from the center of the densest (relatively) cluster (category 3), ensuring visual likelihood. This happens for the following reason. From Eq. 2, we see that for a suitable choice of the coefficients, the function $f_c(\cdot)$ will have a higher score (for all the images in the region), when a chosen import vector (image) in the region is very similar to all the other images in the region. Such a choice would contribute significantly to the minimization of Eq. 3, since many terms in the first part of Eq. 3 can be simultaneously minimized. Then, in a sequential greedy selection process, the next best improvement is achieved by selecting an import vector from another dense cluster, and so on. Thus, the first 4 import vectors are placed nearly at the centers of the 4 clusters. As a result, our formulation naturally maintains a good trade-off between visual likelihood and diversity. Moreover, since we expect similar images to belong to the same classes, semantic likelihood is also satisfied. Also, the number of import vectors that come from each cluster is dependent on the relative size of each cluster with respect to other clusters. Therefore, visual balancing takes place automatically through our objective function. Since this also implies balanced selection of import vectors at the class level, semantic balancing is ensured. See for example, the cluster in category 3 is very big resulting in more vectors being selected from that category.

C. Distribution Regularization

In many practical scenarios, it is important to have flexibility in selecting the subset according to some user defined requirements. For example, while preparing a slide show for a web page discussing *political* news related to Oil Spill, it is more appropriate to include more images that represent *political* aspects than others. In practice, such a scenario requires the generated summary to meet some specified requirement with respect to a category distribution. This can be done as follows. Let $\mathbf{P}_t$ denote a target distribution over the categories that the summary should satisfy; that is, we want $P_t(j) = \frac{1}{|\mathbf{S}|}\sum_{i\in\mathbf{S}} \mathbf{q}_{i,j}, \forall j = 1,\ldots,M$. Note that the right hand side of this equation is nothing but the class distribution of the summary. We *add* a KL-divergence based regularization term $\eta\text{KL}(\mathbf{P}_t, \frac{1}{|\mathbf{S}|}\sum_{i\in S}\mathbf{q}_i)$ to the objective function (Eq. 3); this term minimizes the difference between the target distribution $\mathbf{P}_t$ and distribution of the summary. Here, η controls the contribution of this term. Thus, minimization of Eq. 3 with the KL term decreases the difference between the target distribution and the distribution of the summary. Note that the newly added KL-divergence term is *not dependent* on the classifier coefficients $\mathbf{a}$. Therefore, optimization of $\mathbf{a}$ in Eq. 3 remains the same as before.

D. Implementation Details

We use a Convolutional Neural Network(CNN) to represent each image I_i as a feature vector $\mathbf{x}_i$ in a low-dimensional space. A three-layer DBN described in [11] gives a 1024-dimensional image feature vector ($\mathbf{x}$) for each image.We use bag-of-words to represent the text (metadata). The raw text data is cleaned using standard techniques like removal of stop-words, words with very low and high frequency, etc. In order to get the topic distribution for each image, we perform Latent Dirichlet Analysis (LDA) [9] on the bag of words feature representation of $\mathcal{T}$. An image is expected to belong to at most $2-3$ categories. Therefore, we set the LDA hyper-parameter α in such a way that LDA gives a sparse distribution of topics per image. We use $\alpha = 0.1$ and $\beta = 0.01$ in all our experiments.

We use the Gaussian kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2})$. The parameter σ controls the influence of the import vectors over the feature space. With the CNN features, the value of σ in the range $[0.5, 1.5]$ works well. We also found that setting the regularization parameter λ to a value in the range $[0.0001, 0.001]$ works well.

III. Related Work

For scene summarization and image browsing applications, Simon et al. [1] propose a clustering based approach to find a set of representative images. Berg et al. [5] addresses the problem of finding a set of iconic images from an image collection of a given object category. However, they ([1], [5]) do not use any text information. Simon et al. [1] use metadata information only to tag the canonical views at the cluster level.

Table I
Datasets

Query Type	No. of Queries	Example	Avg. Size
Country	9	China	$\approx$ 1000
Celebrities	11	Paris Hilton	$\approx$ 1000
Sports	6	Cricket	$\approx$ 1000
Events	6	Oscar	$\approx$ 1000
News	13	-	$\approx$ 1000
Abstract	8	Fashion	$\approx$ 1000
Travel	1	San Francisco	4602

Other work [7], [3], [8] use metadata like tags and geo-tags (location where the image was captured) along with the image features. Approaches presented in [7], [8] work mainly with images of landmarks or world photos. Therefore, they are restrictive from application viewpoint. Fan et al. [3] perform latent semantic analysis of tag information to identify the most significant topics and finds the representative set for topic separately through clustering.

There exists a large amount of work (see [4] and references therein) related to image ranking and providing image search results with images of high relevance, likelihood and diversity for a given query. Eva et al. [4] rank a collection of images using a likelihood score computed for each image in a feature space comprising of both image and text features. Diversity is achieved by not selecting the images with high likelihood in the same region of the feature space. However, it is not clear how much importance should be given to textual versus image features.

IV. Experimental Evaluation

In this section, we demonstrate the effectiveness of our method on several real-life concepts and make comparison with three state-of-the-art methods [1], [3], [4].

A. Experimental Setup

Datasets: We construct 54 datasets ranging over 7 broad concept types and collected from various sources such as Flickr, Twitter, Yahoo! News, etc. Table I shows these concept types along with the number of datasets, sample concepts and average dataset size. For News, the collection is obtained from various sources linked via twitter (e.g., http://twitter.com/cnnbrk). The concept types chosen are very diverse in nature and cover significant fraction of popular queries to any search engine.

Methods for Comparison:

(1) **Eva et al. [4]**: We computed a likelihood score for each image in the joint feature space comprising of image and text features, and selected the images with high likelihood. To ensure diversity, we selected the images sequentially and made the likelihood score in the neighborhood of the selected images very small.

(2) **Simon et al. [1]**: We implemented the clustering based

Table II
METRICS USED FOR DIFFERENT PROPERTIES.

Visual Likelihood (VL)	$-\sum_{j \notin \mathbf{S}} \max_{i \in \mathbf{S}} k(\mathbf{x}_i, \mathbf{x}_j)$
Visual Diversity (VD)	$\max_{i,j \in \mathbf{S},\ i \neq j} k(\mathbf{x}_i, \mathbf{x}_j)$
Semantic Likelihood (SL)	$-\sum_{j \notin \mathbf{S}} \max_{i \in \mathbf{S}} \mathrm{KL}(\mathbf{q}_i, \mathbf{q}_j)$
Semantic Diversity (SD)	$\max_{i,j \in \mathbf{S},\ i \neq j} \mathrm{KL}(\mathbf{q}_i, \mathbf{q}_j)$
Semantic Balance (SB)	$\mathrm{KL}(\mathbf{Q_D}, \mathbf{Q_S})$ $\mathbf{Q_D} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{q_i}, \mathbf{Q_S} = \frac{1}{\lvert\mathbf{S}\rvert}\sum_{i \in \mathbf{S}} \mathbf{q_i}$

method [1] that maximizes the visual likelihood. We did not use any text features.

(3) **Spectral Clustering**: We assigned each image i to the topic having the maximum score in $\mathbf{q}_i$. For each topic, we did spectral clustering on images belonging to it. We fixed the number of clusters per topic to be the number of images to be picked from it. Given a concept summary size L, we found the number of images to be picked for each topic using the average topic distribution of the full collection. Finally, we selected the most representative image from each cluster. Note that this method is similar to that of Fan et al. [3].

EVALUATION METRICS: We compare the various methods both qualitatively and quantitatively. For quantitative evaluation, we propose the metrics given in Table II. These metrics are useful to compare the methods by ranking their performance on a large collection of concept summaries as manual evaluation becomes tedious. We define these metrics using some similarity scores computed between pairs of images in the collection; note that lower values mean better performance. For visual and semantic aspects, we use the kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ and KL divergence respectively. To make a fair comparison, we used same image and text features, and kernel similarity measures for generation of slide shows and evaluation of all the methods.

B. Experimental Results

Qualitative Evaluation

SUMMARY FROM OUR METHOD: Fig. 1(a) shows the summary of 9 images obtained from our method without any category distribution constraint. Note that this corresponds to only 2.1% of the whole dataset. Each row represents a different semantic aspect. Our method is able to cover all important aspects. Observe the visual diversity present in the results. The reader is encouraged to refer to the electronic version for better visual clarity.

COMPARISON WITH OTHER METHODS: The method of Eva et al. [4] (Fig. 1b) picks several images with similar titles from aspect (A). Furthermore, it does not have any image related to the *protests* (D) and *corporate* (C) aspects. For spectral clustering (Fig. 1c), the 25^{th} summary image is same as the 5^{th} image. This is because these images had different descriptions (and hence different semantic aspects). Note that in the method of Fan et al. [3], an image can be assigned to multiple topics. In such a case, picking images individually from each topic cluster can potentially hamper visual diversity. Since Simon et al. [1] (Fig. 1d) do not use metadata, their method also misses out on *protests* (D) and *politics* (E) aspects. Note that the spectral clustering method also misses the topic *protests* (D) as there are few images in that topics, and its proportion to the summary size is very small.

Figure 3. Distribution Regularization: Corporate and Politics

SUMMARY WITH DISTRIBUTION REGULARIZATION: Next, we present the summary generated by our method when we included the class distribution regularization term in the objective function (Section II-C). Fig. 3 shows the summary with higher bias towards the categories (C) and (E) (*Corporate and Political*). We set the distribution parameter η to be 0.5. Note that in contrast to Fig. 1a, the images of visiting politicians (E), CEO visiting (C) and workers (C) dominate the summary. None of the other methods currently support this requirement of specifying distribution constraints.

Quantitative Comparison

We computed the metrics in Table II for all the 54 queries (Table I), and methods discussed in Section IV-A. We ranked the methods for each query on each metric. We also computed an *overall* rank by averaging over the ranks obtained for the 5 metrics. Finally, we average these ranks over the queries. To get more insight, we computed these averages for each query type separately, and also over *all* the queries. Due to space limitation, we show average overall ranking results only for three query types, namely, SPORTS, EVENTS and CELEBRITIES, and over all the queries; see Fig. 4. All the results are shown as a function of summary size (specified in terms of percentage of dataset size). Since Simon et al. [1] do not use textual metadata, it is unfair to compare their method on semantic metrics; therefore, we exclude their method from comparison on overall ranking.

OVERALL RANKING (Figs. 4(a-d)): The results clearly

show superior performance of the proposed method on the SPORTS and EVENTS query types (Figs. 4(a,b)). Similar performance was observed on the COUNTRY query type. On the CELEBRITIES query type (Fig. 4(c)), the performance of all the methods is same for large summary sizes; our method gives better performance for smaller summary sizes. Similar performance was observed on ABSTRACT and NEWS query types. The ranking over all the queries is shown in Fig. 4(d); clearly, the proposed method performs better than the spectral clustering and Eva et al. methods. We observed that the performance difference is even more significant when the overall ranking is computed only on the SPORTS, EVENTS and COUNTRY query types.

To understand the above mentioned performance behaviors, we analyzed the correlation between the model predicted class distribution obtained using only image features and the corresponding topic distribution. We observed that our method gave significantly better performance when this correlation is high. For example, the correlation values were around 0.5 for the SPORTS, EVENTS and COUNTRY query types, and it was around 0.25 for remaining query types. Note that *low correlation* on some CELEBRITIES queries such as *Angelina Jolie* is expected; for example, it is difficult to associate her images with different *award ceremonies* using only image features. On the other hand, for concepts like SPORTS, we found strong correlation between text and image features (e.g., different teams have their own uniform colors). We also noticed that for News with *twitter* as the source (Table I), poor correlation was due to inferior performance of LDA; it is known that LDA does not perform so well on short text documents as available from *twitter*.

OVERALL PERFORMANCE: Our method (black curve) (Fig. 4(d)) does well on *all aspects* for summary size of $< 5\%$. Note that for any reasonable dataset size, the desired summary size is often much less than 5% and the performance becomes critically important for large datasets (e.g., *San Francisco*), where summary size is very small.

EFFECT OF NUMBER OF TOPICS(M): We observed that as the number of topics M decreased, the performance difference between the methods decreased for large summary sizes. Note that the performance becomes predominantly controlled by the image features as M becomes small; hence, there was some degradation in the performance. However, our method still performed better when the summary size was small.

V. CONCLUSION

In this paper we proposed a novel classification based framework for concept summarization. Our framework is applicable for a wide range of concepts. Based on our extensive experimental results on various important concept types, we find that our method is well-suited and outperforms other methods under the following conditions: (1) the summary size is small, (2) the number of LDA topics is not very small, and (3) there is correlation between the semantic aspects in the textual description and image features. The performance our method is comparable to other methods otherwise.

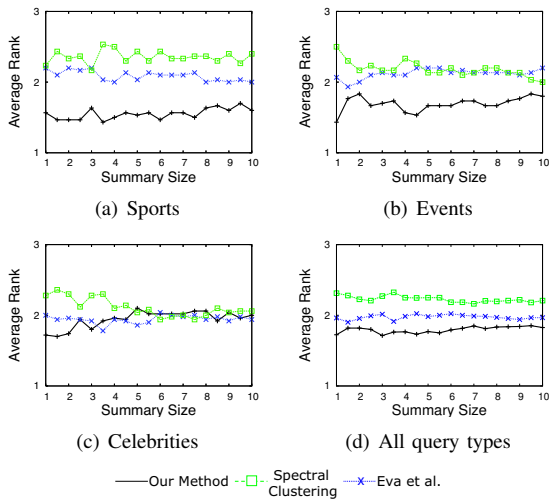

Figure 4. Comparison of different methods: average rank is computed by averaging over all the 5 metrics followed by averaging over the queries.

REFERENCES

[1] I. Simon, N. Snavely, and S. M. Seitz, "Scene summarization for online image collections," in *ICCV*, 2007.

[2] L. Li, K. Zhou, G.-R. Xue, H. Zha, and Y. Yu, "Enhancing diversity, coverage and balance for summarization through structure learning," in *WWW*, 2009.

[3] J. Fan, Y. Gao, H. Luo, D. Keim, and Z. Li, "A novel approach to enable semantic and visual image summarization for exploratory image search," in *MIR*, 2008.

[4] E. Hörster, M. Slaney, M. Ranzato, and K. Weinberger, "Unsupervised image ranking," in *LS-MMRM*, 2009.

[5] T. Berg and A. Berg, "Finding iconic images," *CVPR Workshop*, 2009.

[6] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking via information bottleneck principle," in *MULTIMEDIA*, 2006.

[7] L. S. Kennedy and M. Naaman, "Generating diverse and representative image search results for landmarks," in *WWW*, 2008.

[8] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *WWW*, 2009.

[9] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[10] J. Zhu and T. Hastie, "Kernel logistic regression and the import vector machine," *Journal of Computational and Graphical Statistics*, vol. 14, pp. 185–205, 2005.

[11] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," *CVPR*, June 2007.