Learning to Summarize Web Image and Text Mutually

Piji Li, Jun Ma and Shuai Gao
School of Computer Science & Technology,
Shandong University, Jinan, 250101, China
lipiji.sdu@gmail.com
majun@sdu.edu.cn
gao_shuai@mail.sdu.edu.cn

ABSTRACT

We consider the problem of learning to summarize images by text and visualize text utilizing images, which we call *Mutual-Summarization*. We divide the web image-text data space into three subspaces, namely pure image space (PIS), pure text space (PTS) and image-text joint space (ITJS). Naturally, we treat the ITJS as a knowledge base.

For summarizing images by sentence issue, we map images from PIS to ITJS via image classification models and use *text summarization* on the corresponding texts in ITJS to summarize images. For text visualization problem, we map texts from PTS to ITJS via text categorization models and generate the visualization by choosing the semantic related images from ITJS, where the selected images are ranked by their confidence. In above approaches images are represented by color histograms, dense visual words and feature descriptors at different levels of spatial pyramid; and the texts are generated according to the *Latent Dirichlet Allocation* (*LDA*) topic model. *Multiple Kernel* (*MK*) methodologies are used to learn classifiers for image and text respectively. We show the Mutual-Summarization results on our newly collected dataset of *six big events* ("Gulf Oil Spill", "Haiti Earthquake", etc.) as well as demonstrate improved *cross-media retrieval* performance over existing methods in terms of *MAP*, *Precision* and *Recall*.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval—*retrieval models*

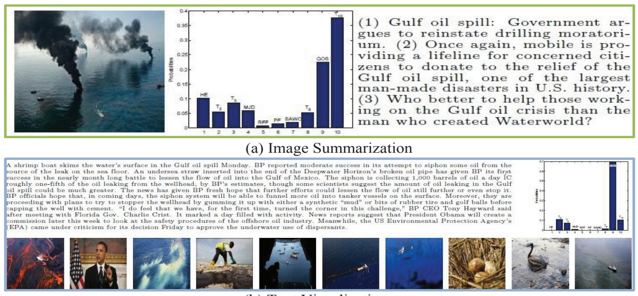
General Terms

Algorithms, Experimentation.

Keywords

Mutual-Summarization, image-text joint space, topic model, cross-media retrieval, multiple kernel learning

ICMR '12, June 5-8, Hong Kong, China



(a) Image Summarization

(b) Text Visualization

Figure 1: Illustration of the Mutual-Summarization results for "Gulf Oil Spill".

1. INTRODUCTION

For a pure image without any text information as shown in left of Figure 1(a), how to generate a set of high level semantic sentences to describe the events happening in this still image (e.g., "Gulf Oil Spill")? For a long news article or some short sentences as shown in Figure 1, how to give a visual display using some existing web images? To address these problems, we propose a framework called "*Mutual-Summarization*". Our work targets improving the performance of some *Computer Vision* and *Information Retrieval* problems, such as image classification, image annotation and description using sentences, cross-modal multimedia retrieval, etc.

Over the last decade there has been a massive explosion of multimedia content on the web. We concentrate on documents containing images and text, although many of the ideas would be applicable to other modalities. It is evident that the web image-text data space could be divided into three sub-spaces:

Space I: pure image space (PIS). Images in this space are all of a single image without semantic text information. Some images in PIS are shown in Figure 1.

Space II: pure text space (PTS). Text documents in this space have no images embedded in them. Some text in PTS are shown in Figure 1.

Space III: image-text joint space (ITJS). With the ongoing explosion of Web-based multimedia content, it is possible and convenient to collect large datasets containing richer image-text data. Examples include news archives, or Wikipedia pages, where images are related to complete long text articles, not just a few tags and short sentences. These rich multimedia information could be used to address many difficult problems as a knowledge base, such as computer vision [18] and cross-modal multimedia retrieval [25].

Based on this partition of image-text data space, the Mutual-Summarization problem can be tackled by utilizing two procedures: *Image Summarization* and *Text Visualization*.

Our contributions include: we introduce a dataset containing six big events ["Gulf Oil Spill (GOS)", "Haiti Earthquake (HE)", "Michael Jackson Died (MJD)", "Pakistan Floods (PF)", "Russian Forest Fires (RFF)" and "South Africa World Cup (SAWC)"]. This dataset on six events is treated as an important knowledge base for our framework. In image summarization procedure, we map images from PIS to ITJS via image classification model and describe these images utilizing several high level semantic sentences. These sentences are summarization of text, generated via the MEAD text summarizer [24]. For text visualization procedure, we map text from PTS to ITJS via text categorization model and then give a visual display utilizing images with high confidences in ITJS. The images are represented as color histograms, distribution of edges, dense visual words and feature descriptors at different levels of spatial pyramid [17]. The text is represented as a sample from a hidden topic model, learned with latent Dirichlet allocation [4]. We employ Multiple Kernel SVM (MK-SVM) [8, 26], Multiple Kernel KNN (MK-KNN) and Semantic Correlation Matching (SCA) [25] to learn classifiers for images and text respectively.

The rest of this paper is organized as follows. It starts with a brief review of related works in Section 2 while the Mutual-Summarization framework is proposed in Section 3. The experimental results and discussions are provided in Section 4. Concluding remarks and future work directions are listed in Section 5.

2. PREVIOUS WORK

Web images summarization component is extremely important and also the most difficult problem in our framework. There are several related studies of literature on this problem, such as action and event classification (in image space), sentence generation for still images. Moreover, the Mutual-Summarization problem can be treated as a model of *Cross-Media Retrieval* and some related studies are also be introduced.

2.1 Events in Images

For the purpose of describing what is happening in a still image, researchers in the field of *Computer Vision* have done some exploratory work in the last five years: from event classification to sentence generation.

Event classification in still images has not been widely studied with the exception of few related papers focused on specific domains. [10] discuss a generative model approach for classifying complex human activities given a single static image in a graphical model representation. [8] investigates more generic recognition methods with bag-of-features and part-based representations for recognizing human actions in still images. There are few attempts to generate sentences and summarization from visual data. [13] generates sentences narrating a sports event in video using a compositional model based around AND-OR graphs. The relatively stylised structure of the events helps sentence generation. [29] presents an more sophisticated image parsing to text description (I2T) framework that generates text descriptions of image and video content based on image understanding from a complex database. [9] describes a system that compute a score linking an image to some manually annotated sentence. These methods generate a direct representation of what objects exist and what is happening in a scene, and then decode it into a sentence. In other words, the sentence generation systems are built on top of the output of multiple-objects recognition systems. However, it has been difficult to establish the value of object recognition for event sentence generation in this cascade manner, mainly because object recognition is still a largely unsolved problem and there will be many objects in an image. Therefore, it is questionable whether the output of any object recognition algorithm is reliable enough to be directly used for event sentence generation.

We focus on the problem of summarizing images using high-level semantic sentences or short articles collected from the Internet, not just describing "what are there" or "what is happening" in images.

2.2 Cross-Media Retrieval

The first generation of cross-modal systems originate from the research on the problem of automatic extraction of semantic descriptors from images [2, 5, 12, 15], which support text-based queries of image databases that do not contain text metadata. However, images are simply associated with keywords, or class labels, and there is no explicit modeling of free-form text. Some notable exceptions are the work of [3], where separates "latent-space" models are learned for images and text, in a form suitable for cross-media image annotation and retrieval. In parallel, advances have been reported in the area of multi-modal retrieval systems. These are extensions of the classic single-modal systems, where a single retrieval model is applied to information from various modalities. This can be done by fusing features from different modalities into a single vector [22, 28], or by learning different models for different modalities and fusing their outputs [16, 27]. However, most of these approaches require multi-modal queries, queries composed of both image and text features. An alternative paradigm is to improve the models of one modality (say image) using information from other modalities (e.g., image captions) [20, 23]. Lastly, it is possible to design multi-modal systems by mapping images and text to a same space and correlations between the two components are learned. Then the cross-modal document retrieval could be solved via retrieving the text that most closely matches a query image, or retrieving the images that most closely match a query text [25].

We focus on the problem learning to summarize images with text and display text with images for some big events based on the dataset collected from the Internet. Naturally, the Mutual-Summarization results could improve the *Cross-Media Retrieval* performance.

3. MUTUAL-SUMMARIZATION

In this section, we present the approach of learning to mutually summarize web image and text. We introduce the image summarization procedure and the text visualization procedure respectively.

3.1 Image Summarization

For a set of pure images $\mathcal{I} = \{I_1, I_2, \cdots I_{|I|}\}$ in $\Re^I$ and a set of sentences $\mathcal{S} = \{S_1, S_2, \cdots S_{|S|}\}$ in $\Re^S$, whenever the image and text data spaces $\Re^I$ and $\Re^S$ have a natural correspondence, image summarization reduces to a classical

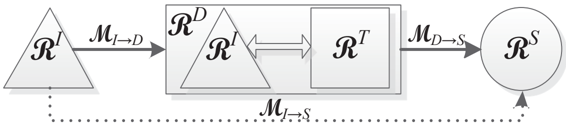


Figure 2: Illustration of learning to summarize images using text summarization.

retrieval problem as shown in the dotted line of Figure 2. Let

$$\mathcal{M}_{I\to S} : \Re^I \to \Re^S \tag{1}$$

be an invertible mapping between the two spaces, where $\mathcal{M}_{I\to S}$ denotes the mapping from $\Re^I$ to $\Re^S$. Given an image $I_i \in \Re^I$, it suffices to find the nearest neighbor to (I_i) in $\Re^S$. In this case, the summarization problem reduces to the design of an effective similarity function for the determination of nearest neighbors.

While images and text are different objects and different representations tend to be adopted for images and text, there is typically no natural correspondence between $\Re^I$ and $\Re^S$. We employ an indirect approach to map images in $\Re^I$ to summarization sentences in $\Re^S$. Following the illustration in Figure 2, we split the mapping $\mathcal{M}_{I\to S}$ into three sub-mappings:

$$\begin{aligned}\mathcal{M}_{I\to S} &\approx \mathcal{M}_{I\to D} + \mathcal{M}_{D\to S} \\ &\approx \mathcal{M}_{I\to I_D} + \mathcal{M}_{I_D\leftrightarrow T_D} + \mathcal{M}_{T_D\to S}\end{aligned} \tag{2}$$

We define $\mathcal{D} = \{D_1, D_2, \cdots D_{|D|}\} \in \Re^D$ as the image-text documents in the image-text joint space (ITJS). We let

$$D_i = \langle I_i, T_i\rangle \tag{3}$$

as *image-text* pair document. We make an important assumption here: given an image-text document $D_i \in \mathcal{D}$ in $\Re^D$, I_i and T_i is a semantic relevant pair, i.e., I_i is semantic relevant to T_i, and vice versa. Based on this assumption and our knowledge base of image-text documents $\mathcal{D}$, we omit the learning procedure of $\mathcal{M}_{I_D\leftrightarrow T_D}$.

Therefore, the remaining work is to build mapping $\mathcal{M}_{I\to I_D}$ and $\mathcal{M}_{T_D\to S}$. We reduce this two problems to image classification and automatic text summarization problems.

3.1.1 Automatic Summarization

We employ MEAD [24] to generate summarization for text. MEAD is a publicly available toolkit for multi-lingual summarization and evaluation. The toolkit implements multiple summarization algorithms (at arbitrary compression rates) such as position-based, Centroid, TF*IDF, and query-based methods. MEAD can perform many different summarization tasks. It can summarize individual documents or clusters of related documents (multi-document summarization). MEAD includes two baseline summarizers: lead-based and random. Lead-based summaries are produced by selecting the first sentence of each document, then the second sentence of each, etc. until the desired summary size is met. A random summary consists of enough randomly selected sentences (from the cluster) to produce a summary of the desired size.

We utilize lead-based individual documents MEAD summarizer to map text in $\Re^{T_D}$ to sentences in $\Re^S$, and the compression percentage we used is 25%:

$$\mathcal{M}_{T_D\to S} : \Re^{T_D} \to \Re^S \tag{4}$$

3.1.2 Image Classification

For the purpose of mapping image $I_i \in \Re^I$ to image $I_j \in \Re^{I_D}$ using $\mathcal{M}_{I\to I_D}$,

$$\mathcal{M}_{I\to I_D} : \Re^I \to \Re^{I_D} \tag{5}$$

we reduce this problem to 6-class image classification task. Given a set of N training examples $\{(I^{(n)}, Y^{(n)})\}_{n=1}^N$, we learn a discriminative and efficient classification function $H : \mathcal{I} \times \mathcal{Y} \to \mathcal{R}$ over an image I and its class label Y, where $\mathcal{I}$ denote the input space of images and $\mathcal{Y} = \{1, 2, i, i+1, \cdots, |C|\}$ is the set of class labels, here $|C| = 6$. H is parameterized by Θ. For a new pure image $I_i \in \Re^I$, we map I_i to the six events semantic space via $H(I_i, Y; \Theta)$:

$$Y^* = \arg\max_{Y\in\mathcal{Y}} H(I_i, Y; \Theta) \tag{6}$$

Thereafter, we map $I_i \in \Re^I$ to some nearest $I_j \in \Re^{I_D}$ in event class Y^*. The mapping $\mathcal{M}_{I\to I_D}$ is built.

We mainly employ Multiple Kernel SVM (MK-SVM) [8, 26] to learn the mapping $\mathcal{M}_{I\to I_D}$ from knowledge base we collect, comparing with Multiple Kernel KNN (MK-KNN) and Semantic Correlation Matching (SCA) [25].

(a) Multiple Kernel SVM (MK-SVM)

The first method to learn the mapping $\mathcal{M}_{I\to I_D}$ is Multiple Kernel SVM [8, 26]. In implements, the function $H(I_i, Y; \Theta)$ is learnt, along with the optimal combination of state-of-art features and spatial pyramid levels, by using the MKL technique. The function $H(I_i, Y; \Theta)$ is the discriminant function of a Support Vector Machine (SVM), and is expressed as

$$H(I, Y; \Theta) = \sum_{i=1}^{N} \theta_i [K(\varphi(I), \varphi(I^i)), Y_i] \tag{7}$$

where $\varphi(I^i), i = 1, 2\cdots N$ denote the feature descriptors of N training images, $Y_i \in \mathcal{Y}$ is their class labels, and K is a positive definite kernel, obtained as a liner combination of histogram kernels by η:

$$\begin{aligned}K(\varphi(I), \varphi(I^i)) &= \sum_{k=1}^{\#\varphi} \eta_k k(\varphi_k(I), \varphi_k(I^i)) \\ \sum_{k=1}^{\#\varphi} \eta_k &= 1\end{aligned} \tag{8}$$

where $\#\varphi$ is the number of features to describe the appearance of images. For example, $\#\varphi = 2$ for two kinds of features (Color Histogram and Pyramid SIFT). MKL learns both the coefficient θ_i and the histogram combination weights $\eta_k \in [0, 1]$.

We consider three types of kernels, which are different in their discriminative power and computational cost. Our gold standard is histogram intersection kernel of the form

$$k(x, y) = \sum_{i=1}^{n} \min(x_i, y_i) \tag{9}$$

We also consider radial basis function (RBF) kernel and linear kernel to compare the performance.

(b) Multiple Kernel KNN (MK-KNN)

K-nearest neighbors algorithm (KNN)[1] is a method for classifying objects based on closest training examples in the

[1] http://en.wikipedia.org/wiki/KNN

feature space. Similarity metric is the most important component of KNN. We employ the combination of multiple kernels (see Eq.(8)) as the similarity metric $s(x, y)$ of MK-KNN:

$$s(x, y) = K(x, y) \tag{10}$$

where x and y denote visual feature histograms of two images.

(c) Semantic Correlation Matching (SCM)

Nikhil Rasiwasia [25] utilizes Canonical correlation analysis (CCA) to learn a basis of canonical components for images and text respectively, i.e., directions $w_i \in \Re^I$ and $w_t \in \Re^T$ along which the data is maximally correlated, i.e.,

$$\max_{w_i \neq 0, w_t \neq 0} \frac{w_i^T \sum_{IT} w_t}{\sqrt{w_i^T \sum_{II} w_i}\sqrt{w_t^T \sum_{TT} w_t}} \tag{11}$$

After the optimization of (11) being solved, images and text can be mapped to a same subspace $\mathcal{U}$ based on w_i and w_t. Semantic Correlation Matching (SCM) is built via multi-class logistic regression, thereafter images and text are mapped to a semantic space $\mathcal{S}$.

In the classification progress, we employ the semantic presentation $SCM(I) \in \mathcal{S}$ instead of $\varphi(I) \in \Re^I$. Finally we employ the multi-class SVM to learn the mapping $\mathcal{M}_{I \to I_D}$.

3.1.3 Sentence Selection

When finishing the image classification procedure, a new pure image $I_i \in \Re^I$ can be mapped to $\Re^{I_D}$. According to $\mathcal{M}_{I_D \leftrightarrow T_D}$ and $\mathcal{M}_{T_D \leftrightarrow S}$, a list of sentences $S \in \Re^S$ are selected to summarize I_i, ranked by their confidence with I_i. We select the combination of multiple kernels (see Eq.(8)) as the confidence function $Conf(x, y)$:

$$Conf(x, y) = K(x, y) \tag{12}$$

For a new pure image $I_i \in \Re^I$ and a sentence $S_j \in \Re^S$, we can not compute the confidence $Conf(I_i, S_j)$ directly. Based on the mapping $\mathcal{M}_{I \to S}$, we can get the approximate semantic confidence by formula (13):

$$Conf(I_i, S_j) \approx Conf(I_i, D_j) \approx Conf(I_i, I_{Dj}) \tag{13}$$

Therefore the confidence between two images can be computed directly for the reason that they are in a same data space PIS. Assume the event class label of image I_i is c_i, we can get the top q images $\{I_{D_1}, I_{D_2}, \cdots, I_{D_{|q|}}\} \in \Re^{I_D}$ in class c_i. According to mapping $\mathcal{M}_{I_D \leftrightarrow T_D}$, top q articles $\{T_{D_1}, T_{D_2}, \cdots, T_{D_{|q|}}\} \in \Re^{T_D}$ are selected to describe image I_i. For convenience, we employed MEAD [24] automatic text summarizer to extract the most important sentence for each $T_{D_{|i|}}$. Finally, q sentences are selected to summarize the semantic information of image I_i. In experiments, we let $q = 3$.

3.2 Text Visualization

3.2.1 Mappings

Learning to summarize pure text using web images, which also called text visualization, is to map pure text $T_i \in \Re^T$ to images $I_j \in \Re^I$,

$$\mathcal{M}_{T \to I} : \Re^T \to \Re^I \tag{14}$$

as the dotted line of Figure 3 shows. While images and text are different objects and different representations tend to be adopted for images and text, there is typically no natural correspondence between $\Re^T$ and $\Re^I$.

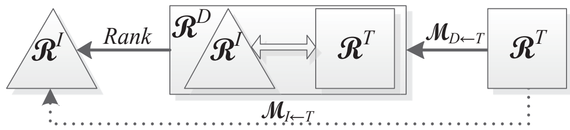


Figure 3: Illustration of learning to visualize text using images.

As the method we proposed in Section 3.1, following the illustration in Figure 3, we split the $\mathcal{M}_{T \to I}$ procedure into two sub-mappings:

$$\begin{aligned} \mathcal{M}_{T \to I} &\approx \mathcal{M}_{T \to D} \\ &\approx \mathcal{M}_{T \to T_D} + \mathcal{M}_{T_D \leftrightarrow I_D} \end{aligned} \tag{15}$$

Since the mapping $\mathcal{M}_{T_D \leftrightarrow I_D}$ has been automatically built based on our image-text knowledge base, we just concern the mapping $\mathcal{M}_{T \to T_D}$,

$$\mathcal{M}_{T \to T_D} : \Re^T \to \Re^{T_D} \tag{16}$$

We also reduce the map procedure to a multi-class text categorization problem. The representation of text in $\Re^T$ is derived from the *latent Dirichlet allocation* (LDA) model [4]. LDA is a generative model for a text corpus, where the semantic content of a text is summarized as a mixture of topics. More precisely, a text is modeled as a multinomial distribution over K topics, each of which is in turn modeled as a multinomial distribution over words. Each word in a text T_i is generated by first sampling a topic z from the text-specific topic distribution, and then sampling a word from that topic's multinomial. In R^T text documents are represented by their K-dimension topic assignment probability distributions [25].

Similarly, we employ multi-class SVM, KNN and Semantic Correlation Matching (SCM) to implement text categorization problem. For SVM and KNN methods, we represented text as the LDA based features. For SCM, we unitize the semantic representation $SCM(T_i) \in \mathcal{S}$ and thereafter employ SVM to learn the mapping $\mathcal{M}_{T \to T_D}$.

3.2.2 Image Selection

When finishing the text categorization procedure, a new pure text $T_i \in \Re^T$ can be mapped to $\Re^{T_D}$. According to $\mathcal{M}_{T_D \leftrightarrow I_D}$, a list of representative images $I \in \Re^{I_D}$ are selected to visualize T_i, ranked by confidence with T_i. We utilize single kernel value $k(x, y)$ as the confidence function, i.e. $Conf(x, y) = k(x, y)$. According to formula (17)

$$Conf(T_i, I_j) \approx Conf(T_i, D_j) \approx Conf(T_i, T_{Dj}) \tag{17}$$

we select the top p images from $\Re^{I_D}$ to visualize text T_i. In experiments, we let $p = 10$.

4. EXPERIMENTS

4.1 Dataset

We collect about 1200 news articles in total for 6 big events: "Gulf Oil Spill (GOS)", "Haiti Earthquake (HE)", "Michael Jackson Died (MJD)", "Pakistan Floods (PF)", "Russian Forest Fires (RFF)" and "South Africa World Cup (SAWC)". Each article contains at least one image embedded into the text. Thereafter, the dataset was pruned by removing the unwanted images to ensure that each text contains only one image. The final corpus contains a total of 1200 image-text pairs, annotated with a label from the 6 events classes as shown in Figure 1. A random split was used to produce a training set of 800 (67%$\times$ 1200) documents, and a test set of 400 (33%$\times$ 1200) documents. The training set is treated as a knowledge base ($\in \Re^D$). In the image summarization procedure, the left 400 images are treated as the test set ($\in \Re^I$). In the text visualization procedure, the corresponding 400 text are treated as the test set ($\in \Re^T$).

For convenience, we utilize "GOS", "HE", "MJD", "PF", "RFF" and "SAWC" to denote the labels of the six events we collect.

4.2 Image and Text Representation

The text documents are represented by their topic assignment probability distributions via LDA (see Section 3.2).

The descriptors of the appearance of images are constructed from a number of different state-of-the-art features. These are the features used in [7, 11, 17, 26, 19]:Dense SIFT Words (BoW) [17], Histogram of Oriented Edges (HOG) [7],Gist [21], Region Color Histogram (RCH) and Spatial Pyramid [17, 26] (SP-BoW and SP-HOG).

4.3 Image Summarization Results

4.3.1 Kernel Selection

It is significant to select a perfect kernel for the image classification methods MK-SVM, MK-KNN and SCM, which we used in our framework to learn the mapping $\mathcal{M}_{I \to I_D}$ from knowledge base. We run five times five-fold cross validation via multi-class LibSVM of Matlab version [6] and get the mean classification accuracy for each visual feature on each kernel function, the accuracy and time cost are shown in Figure 4.This work is accomplished via Matlab on a PC with two 2.93GHz CPUs.

For accuracy, as shown in Figure 4(a), the histogram intersection kernel (HI-K, see Eq.(9)) outperforms the radial basis function (RBF) and the linear kernel (Linear-K) by 11.11% and 16.27% on average. Moreover, the image representation using SP-BoW outperforms other visual features on image classification problem. For efficiency, as shown in Figure 4(b), HI-K outperforms RBF and Linear-K by 83.06% and 57.61% on average.

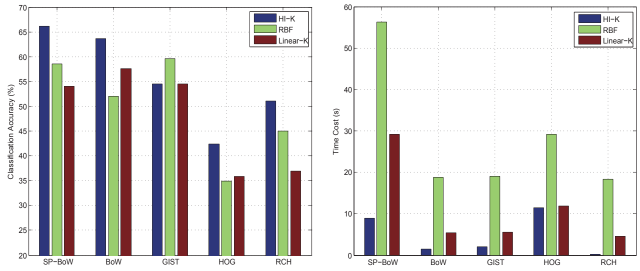
It is evident that histogram intersection kernel (HI-K) is an effective and efficient kernel for image classification problem, which is selected as the base kernel function in our framework.

4.3.2 The Combination Parameters η_k

We learn the optimal combination parameters η_k (weights for SP-BoW, GIST, HOG and RCH) via MK-SVM technique. Firstly, we tune parameters at a coarse-grained level (0.1) to select features. Thereafter, we tune parameters at a fine-grained level (0.01) to search the optimal combination parameters. The optimal coarse-grained tuning results are shown in Table 1.

Table 1: The optimal coarse-grained tuning.

feature	SP-BoW	GIST	HOG	RCH	Accuracy
η_k	0.8	0.0	0.0	0.2	69.70%



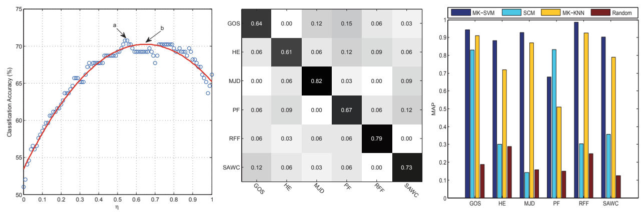
(a) (b)

Figure 4: Accuracy (%) and time cost (seconds) comparison of image classification on different visual features and different kernel functions.

It is interesting that SP-BoW and RCH are selected weighted by 0.8 and 0.2 respectively, while GIST and HOG are omitted. Intuitively, we analyze that the theories of SP-BoW and RCH are completely different and the fusion of them will improve the classification performance.

Assume the combination parameter for SP-BoW is η, and naturally, the parameter for RCH is $(1-\eta)$ according to Eq.(8). Then we tune the parameter η at the fine-grained level (0.01) and the tuning results is shown in Figure 5(a).Point a is the optimal η at the original discrete space and b is the optimal point after least squares fitting. At point a: $\eta = 0.55, accuracy = 70.72\%$; and at point b: $\eta = 0.65, accuracy = 70.22\%$. In experiments we select the parameter at point a. i.e., the weight for SP-BoW is $\eta = 0.55$ and the weight for RCH is $1 - \eta = 0.45$. The final 6-class image classification results based on MK-SVM are shown in Figure 5(b).

4.3.3 Summarization for Images



(a) (b) (c)

Figure 5: (a) The fine-grained tuning of parameters η between SP-BoW and RCH. (b) Confusion matrix of 6-class image classification obtained by MK-SVM. (c) MAP performance of image summarization for the six event categories. (For clarity, you can increase the display rate of this page to 300%.)

Whenever the mapping $\mathcal{M}_{I \to I_D}$ is built, for a pure image $I_i \in \Re^I$, several sentences will be generated to describing the semantic content of I_i. Actually, it is similar with the problem of searching text using images. Therefore, we can evaluate the images summarization results via evaluation standard used in information retrieval.

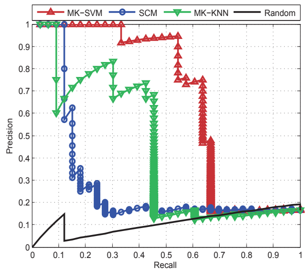
In all cases, performance is measured with precision-recall (PR) curves and mean average precision (MAP) [1]. MAP is obtained as the mean of average precisions over a set of queries. Given a query, its MAP is computed by Eq.(18), where N_{rel} is the number of relevant images, N is the number of total retrieved images, $rel(n)$ is a binary function

Table 2: **Words in each topic of the 6-events dataset.**

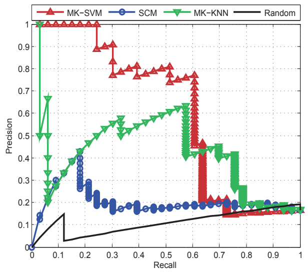
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
haiti	dai	world	jackson	fire	flood	world	chang	oil	includ
earthquak	official	people	michael	russia	pakistan	south	nature	spill	plan
haitian	week	time	death	region	people	cup	caus	gulf	month
people	report	new	pop	forest	on	africa	term	bp	govern
port	time	live	die	moscow	water	team	increas	coast	provid
princ	accord	look	famili	emerg	aid	game	anim	water	help
countriy	move	seen	report	ministri	countriy	soccer	human	mexico	billion
au	continue	life	music	people	affect	african	energi	drill	respons
help	citiy	photo	angel	russian	govern	match	percent	on	cost
school	start	don	lo	burn	relief	stadium	nation	disast	fund

indicating whether the nth image is relevant, and $P(n)$ is the precision at n.

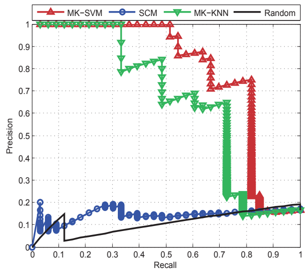
$$MAP = \frac{1}{N_{rel}} \sum_{n=1}^{N} P(n) \times rel(n) \tag{18}$$



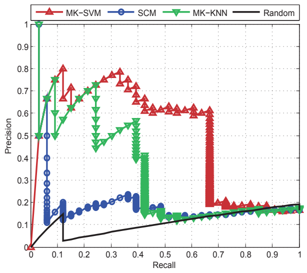
(a) “Gulf Oil Spill”



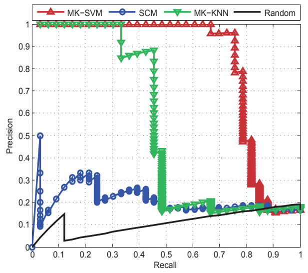
(b) “Haiti Earthquake”



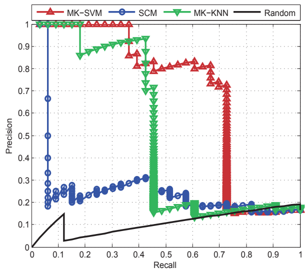
(c) “Michael Jackson Died”



(d) “Pakistan Floods”



(e) “Russian Forest Fires”



(f) “South Africa World Cup”

Figure 6: The precision-recall cures of the image summarization performance for each event category.

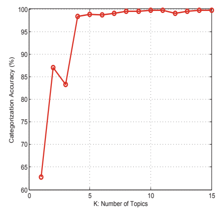
Figure 5(c) shows the MAP performance of image summarization for the six event categories. The average MAP of MK-SVM, MK-KNN and SCM for all six categories are 88.74%, 78.70% and 46.42%. MK-SVM outperforms MK-KNN and SCM by 12.76% and 92.39%.

The precision-recall (PR) curves for each event category are shown in Figure 6. It is evident that MK-SVM performs better than another two methods and MK-KNN is also better than SCM. The reasons for the relatively poor performance of SCM [25] is probably that canonical correlation analysis (CCA) [14] technique can adversely affect classification performance.

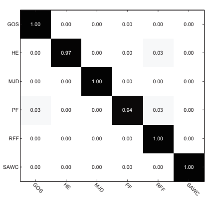
Finally, some image summarization results are displayed in Figure 8. We select three sentences with high confidence to summarize each image.

4.4 Text Visualization Results

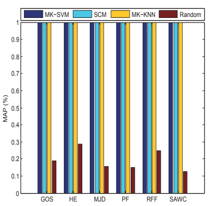
4.4.1 *Text Categorization*



(a)



(b)



(c)

Figure 7: (a) The relation between number of topics K and text categorization performance. (b) Confusion matrix of 6-class text categorization obtained by SVM. (c) MAP performance of text categorization for the six event categories. (For clarity, you can increase the display rate of this page to 300%.)

The pure text in $\Re^T$ are mapped to $\Re^{T_D}$ via a multi-class text categorization problem. In R^T text documents are represented by their K-dimension topic assignment probability distributions via LDA. The number of topics K will effective the performance of text categorization, as shown in Figure 7(a), we select $K = 10$ to get better categorization performance.

Figure 7(b) shows the 6-class text categorization obtained by multi-class SVM. Interestingly, the text dataset we randomly select from the Internet has **strong discriminant power**. The average classification accuracy is 98.48%. The top 10 of most likely words per topic are selected to analyze some properties of the dataset. As shown in Table 2, Topic 1, Topic 4, Topic 5 ,Topic 6, Topic 7 and Topic 9 correspond with the topics of the six big events we collect. Topic 2, Topic 3, Topic 8 and Topic 10 are some latent topics. Since the topics in our dataset are obvious and accurate, we get a sound performance of text categorization. Moreover, the words in each topic can be used to **annotate** or **tag** images.

These annotations and tags are in high-level semantic space, not just describe the objects in images.

4.4.2 Visualization for Text

After mapping $\mathcal{M}_{T \to T_D}$ is built, for a pure text $T_i \in \Re^T$, some images from $\Re^{I_D}$ can be retrieved to visualize T_i, ranked by their confidence. Similarly, this is a retrieval problem and we also employ precision-recall curves and MAP to evaluate the results of text visualization based on SVM, KNN and SCM.

Figure 7(c) shows the MAP for each event category. It is naturally that the MAP of SVM, KNN and SCM are almost 100% because our dataset has strong discriminative power. The same situation also occurs in precision-recall curves for each event, i.e., both SVM and KNN have perfect performance. Moreover, a slightly lower performance is anticipated via SCM.

Finally, some text visualization results are displayed in Figure 9.

5. CONCLUSIONS

We consider the problem of learning to summarize images using text and learning to visualize text using images, which we called Mutual-Summarization. In the future work, we will study new techniques to improve the Mutual-Summarization performance. For instance, the image classification component should be improved via more effective representations and classifiers. Moreover, the performance of automatic text summarization will be studied. Finally, we will extend the knowledge base for more applications.

6. ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of China (60970047,61103151,61173068) and Doctoral Fund of Ministry of Education of China (20110131110028). We wish to thank every body for the contribution of this paper.

7. REFERENCES

[1] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[2] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. Blei, and M. Jordan. Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135, 2003.

[3] D. Blei and M. Jordan. Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 127–134. ACM, 2003.

[4] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 394–410, 2007.

[6] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893. IEEE, 2005.

[8] V. Delaitre, L. I., and S. J. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *British Machine Vision Conference*, 2009.

[9] A. Farhadi, M. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every Picture Tells a Story: Generating Sentences from Images. *ECCV 2010*, pages 15–29, 2010.

[10] L. Fei-Fei and L. Li. What, Where and Who? Telling the Story of an Image by Activity Classification, Scene Recognition and Object Categorization. *Computer Vision*, pages 157–171, 2010.

[11] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[12] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2. IEEE, 2004.

[13] A. Gupta, P. Srinivasan, J. Shi, and L. Davis. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In *IEEE Conference on Computer Vision and Pattern Recognition.*, pages 2012–2019. Citeseer, 2009.

[14] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4):321, 1936.

[15] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 119–126. ACM, 2003.

[16] T. Kliegr, K. Chandramouli, J. Nemrava, V. Svatek, and E. Izquierdo. Combining image captions and visual analysis for image concept classification. In *Proceedings of the 9th International Workshop on Multimedia Data Mining: held in conjunction with the ACM SIGKDD 2008*, pages 8–17. ACM, 2008.

[17] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178. IEEE, 2006.

[18] L. Li and L. Fei-Fei. Optimol: automatic online picture collection via incremental model learning. *International Journal of Computer Vision*, 88(2):147–168, 2010.

[19] P. Li and J. Ma. What is happening in a still picture? In *First Asian Conference on Pattern Recognition (ACPR)*, pages 32–36. IEEE, 2011.

[20] A. Nakagawa, A. Kutics, K. Tanaka, and M. Nakajima. Combining words and object-based visual features in image retrieval. 2003.

[21] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, 155:23–36, 2006.

[22] T. Pham, N. Maillot, J. Lim, and J. Chevallet. Latent semantic fusion model for image retrieval and annotation. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, pages 439–444. ACM, 2007.

[23] A. Quattoni, M. Collins, and T. Darrell. Learning visual representations using images with captions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[24] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang. MEAD - a platform for multidocument multilingual text summarization. In *LREC 2004*, Lisbon, Portugal, May 2004.

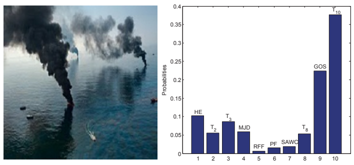
[25] N. Rasiwasia, J. Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. A New Approach to Cross-Modal Multimedia Retrieval. In *Proceedings of ACM International Conference on Multimedia*. ACM, 2010.

[26] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *IEEE International Conference on Computer Vision*, pages 606–613. IEEE, 2010.

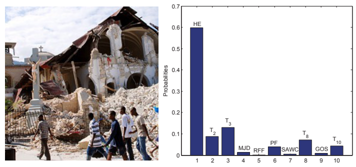
[27] G. Wang, D. Hoiem, and D. Forsyth. Building text features for object image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1374. IEEE, 2009.

[28] T. Westerveld. Probabilistic multimedia retrieval. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 437–438. ACM, 2002.

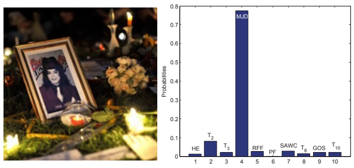
[29] B. Yao, X. Yang, L. Lin, M. Lee, and S. Zhu. I2T: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508, 2010.



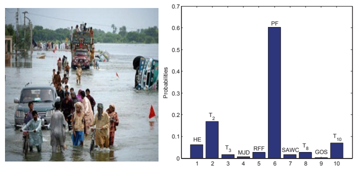
(1) Gulf oil spill: Government argues to reinstate drilling moratorium. (2) Once again, mobile is providing a lifeline for concerned citizens to donate to the relief of the Gulf oil spill, one of the largest man-made disasters in U.S. history. (3) Who better to help those working on the Gulf oil crisis than the man who created Waterworld?



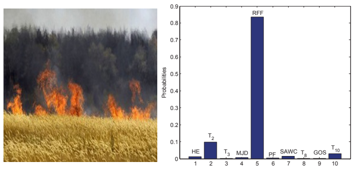
(1) Hundreds of thousands of people have died in Haiti's earthquake , the prime minister told CNN Wednesday. (2) Haitian authorities said the powerful quake destroyed most of the capital city of Port-au-Prince. (3) Haiti's first lady, Elisabeth Debrosse Delatour, reported that "most of Port-au-Prince is destroyed" and that many government buildings had collapsed, Haiti's ambassador to the United States, Raymond Joseph, told CNN Wednesday morning.



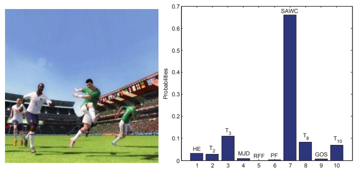
(1) The pop star was rushed to hospital in Los Angeles with suspected cardiac arrest after the star stopped breathing. (2) Pop icon Michael Jackson died following a one hour attempt by a team of emergency physicians and cardiologists to save his life. (3) Michael Jackson's death from a heart attack was concealed for three hours in a bid to change the circumstances of his death, according to claims made in the UK Mirror newspaper.



(1) Officials said they feared further chaos as the water levels of the Sindh and Kabul rivers continue to rise and more rain was expected overnight. (2) A flood survivor carrying relief goods walks past toppled vehicles in Muzaffargarh district, Punjab province, Pakistan. (3) The human and financial toll from devastating floods throughout the subcontinent continued to rise yesterday.



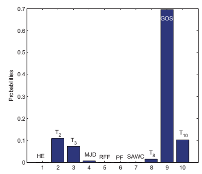
(1) Fierce fire is burning in Russian forests. (2) Hundreds of thousands of firefighters, including army troops, on Saturday battled forest fires raging across central Russia in a heat wave that has killed more than 30 people. (3) Temperatures were forecast to hit 40 degrees Celsius (104 degrees Fahrenheit) in the next few days in several central Russian regions, with the emergency ministry warning of an "extreme risk" of more forest fires.

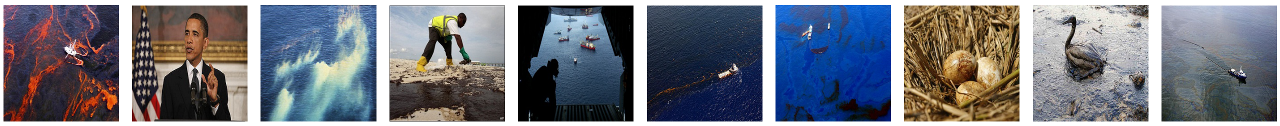


(1) The nation has endured a roller coaster of successes and failures during that time, and earning the right to host the World Cup is one of the great achievements of both Mandela and his extraordinary country. (2) People from across the globe will be travelling to South Africa to watch world cup football and to meet and party with community and cultures from all over the world. (3) The 19th World Cup will be hosted by South Africa in 2010 and will take place between the 11th of June to the 11th of July.

Figure 8: Examples of image summarization results for each event class. Each image is summarized by three sentences with high confidence.

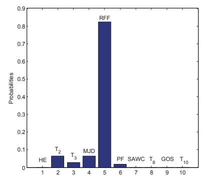
A shrimp boat skims the water's surface in the Gulf oil spill Monday. BP reported moderate success in its attempt to siphon some oil from the source of the leak on the sea floor. An undersea straw inserted into the end of the Deepwater Horizon's broken oil pipe has given BP its first success in the nearly month long battle to lessen the flow of oil into the Gulf of Mexico. The siphon is collecting 1,000 barrels of oil a day ÍC roughly one-fifth of the oil leaking from the wellhead, by BP's estimates, though some scientists suggest the amount of oil leaking in the Gulf oil spill could be much greater. The news has given BP fresh hope that further efforts could lessen the flow of oil still further or even stop it. BP officials hope that, in coming days, the siphon system will be able to funnel more oil into tanker vessels on the surface. Moreover, they are proceeding with plans to try to stopper the wellhead by gumming it up with either a synthetic "mud" or bits of rubber tire and golf balls before capping the well with cement. "I do feel that we have, for the first time, turned the corner in this challenge," BP CEO Tony Hayward said after meeting with Florida Gov. Charlie Crist. It marked a day filled with activity. News reports suggest that President Obama will create a commission later this week to look at the safety procedures of the offshore oil industry. Meanwhile, the US Environmental Protection Agency's (EPA) came under criticism for its decision Friday to approve the underwater use of dispersants.





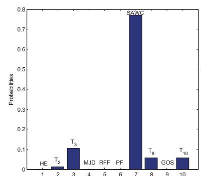
(a) "Gulf Oil Spill"

The death toll from forest fires sweeping across Russia amid a record-breaking heatwave grew to 25 on Friday, with three firefighters among the dead, officials said. The bodies of six residents were discovered in the village of Mokhovoye in the Moscow region, news agencies reported, citing the emergency ministry. The governor of the Ryazan region, one of those worst hit, said that three people had died in the region, in televised comments. A fireman died in hospital from burns after fighting flames on Thursday in a village in the Lipetsk region, the chief doctor at the regional burns centre told the Itar-Tass news agency. The bodies of nine people were found in the Nizhny Novgorod region, the emergency ministry said, updating a provisional toll announced earlier of two. Earlier, the death of a fireman in the Moscow region and five deaths in the Voronezh region were reported. The emergency ministry did not give a total toll for the whole of Russia. Forest fires swept through central Russia amid a record heatwave that has led to droughts in 23 regions and seen the temperature in Moscow hit an all-time record of 38.2 degrees celsius.



(b) "Russian Forest Fires"

This summer all soccer fans will be focused on South Africa for the 2010 World Cup. But, before the tournament begins, gamers can get their paws on EAąfs FIFA World Cup 2010 South Africa and play as any of the 199 qualified teams at all 10 official World Cup stadiums. Gamers can play as their favorite team, run through the tournament, and play in the World Cup Finals to feel the excitement of winning the sportąfs biggest tournament. EA says that everything we love about the World Cup will be reproduced in the game, including the addition of confetti, streamers, and fireworks. What better way to celebrate a World Cup win than with some streamers. You can also play online and take your favorite team through the tournament. If your favorite team didnąft qualify in real life, this is your chance to take your home team through the tournament and into the finals. The game is slated for release on April 27 in North America and April 30 in Asia and Europe on PlayStation 3, Xbox 360, Wii, and PSP.



(c) "South Africa World Cup"

Figure 9: Examples of text visualization results for each event category. Each text is visualized by ten images ranked by their confidence.