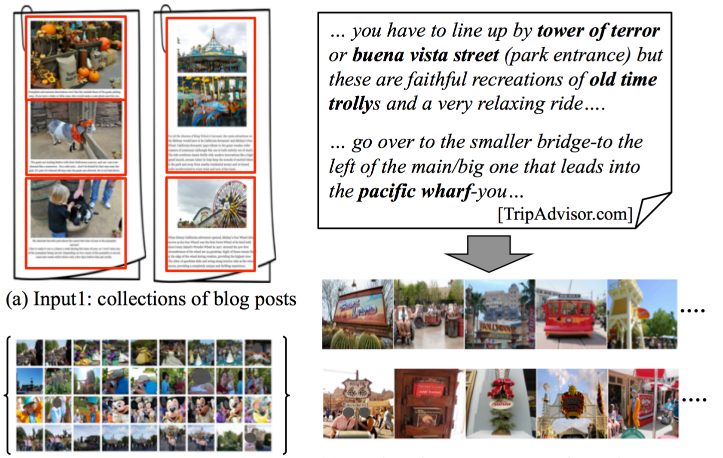
Ranking and Retrieval of Image Sequences from Multiple Paragraph Queries

Gunhee Kim[1], Seungwhan Moon[2] Leonid Sigal[3]
[1]Seoul National University. [2]Carnegie Mellon University. [3]Disney Research Pittsburgh.

Research Objective. We propose a method to rank and retrieve image sequences from a natural language text queries, consisting of multiple sentences or paragraphs. Recently, there has been steady progress toward joint understanding of natural language (NLP) descriptions and contents of images [1, 3, 4]. While most previous work has dealt with the relations between a single sentence and an image or a video, our work extends to the relations between paragraphs and image sequences.

Fig.1 illustrates an intuition of our problem statement with an example of tourism (*e.g.* visiting *Disneyland*). Given a text query consisting of multiple paragraphs, our goal is to automatically retrieve the image sequence that best describe the essence of the query text. Our approach leverages the vast user-generated resource of blog posts and photo streams on the Web. We use blog posts as text-image parallel training data that co-locate informative text with representative images that are carefully selected by users. In order to populate image samples about the topic more densely, we also exploit a large set of photo streams, each of which is a sequence of images that are taken by a single user in one day.

One of our key applications is to visualize visitors' text-only reviews on TRIPADVISOR or YELP, by automatically retrieving the most illustrative image sequences from the Web. Note that general users can understand key concepts and sentiment much easier and quicker with images, as is evident from illustrative travel books. As a problem domain, we focus on a theme park, specifically *Disneyland*, because it is easy to obtain abundant visual and textual data. However, our approach and problem formulation are much broader and are applicable to any domain that has a large set of blog posts with images (more broadly, any mixed image and text media).



*... you have to line up by **tower of terror** or **buena vista street** (park entrance) but these are faithful recreations of **old time trollys** and a very relaxing ride....*

*... go over to the smaller bridge-to the left of the main/big one that leads into the **pacific wharf**-you...*

[TripAdvisor.com]

(a) Input1: collections of blog posts

(b) Input2: sets of photo streams

(c) Retrieve image sequences for a given text

Figure 1: A depiction of our problem statement with *Disneyland* examples. We leverage (a) blog posts to learn the mapping between the sequences of sentences and images, and (b) photo streams to augment the image samples. (c) Our objective is to rank and retrieve image sequences that best describe a given text query consisting of multiple sentences or paragraphs.

Approach. We have three types of input data: (1) A set of visitors' blog posts $\mathcal{B} = \{B^1, \cdots, B^N\}$, which is used as an image-text parallel corpus for training, where N is the number of blogs, (2) A large set of L visitors' photo streams $\mathcal{P} = \{P^1, \cdots, P^L\}$, whose main use is to populate image samples for retrieval, and (3) A set of text-only posts $\mathcal{Q}$ where each post $Q \in \mathcal{Q}$ consists of multiple sentences or paragraphs. We use $\mathcal{Q}$ as a set of text queries.

We formulate our problem as follows. Given a query text $Q \in \mathcal{Q}$, we rank and retrieve a set of image sequences, $\mathcal{S} = \{(S^1, w^1), \cdots, (S^K, w^K)\}$ where S^k is the k-th ranked image sequence, w^k is a ranking score, and K is the number of sets to retrieve. Each S^k consists of images from blog posts $\mathcal{B}$ or photo streams $\mathcal{P}$. In order to represent images and text of input data and uncover the relation between them, we use five syntactic and semantic based text segmentation methods, three text descriptors (Bag-of-Words, TF-IDF, LDA topic distribution), two image descriptors (color SIFT and HOG), and two text-to-image embedding methods (Normalized CCA, KNN).

To learn the best weights of combinations of different model components, we design our ranking and retrieval approach using the structural SVM with latent variables (*e.g.* [2]). We treat text segmentation as a *latent variable*, because correct segmentation in unknown and can be different for different queries. If we denote one segmentation instance by $H \in \mathcal{H}$, the retrieval objective is to find the optimal image sequences S^* for a query Q

$$S^* = \underset{(S,H)\in\mathcal{S}\times\mathcal{H}}{\operatorname{argmax}} \mathcal{F}(Q,S,H) = \underset{(S,H)\in\mathcal{S}\times\mathcal{H}}{\operatorname{argmax}} \alpha \cdot \Phi(Q,S,H) + \beta \cdot \Pi(Q,S,H) \quad (1)$$

where the discriminant function $\mathcal{F}$ is linear in the feature vectors, which are decomposed into two components: The first $\Phi(Q,S,H)$ includes a set of features describing a one-to-one relation between each pair of (s_i, h_i), whereas $\Pi(Q,S,H)$ consists of feature vectors relating S to H as a set.

Experiments. We evaluate the image retrieval performance of our method on a newly collected *Disneyland* dataset, which consists of more than 10K blog posts with 120K associated images, and 6K photo streams of more than 540K unique images. We present comprehensive empirical studies comparing between five text segmentation and three text description methods and their combinations. Our approach using latent structural SVM can efficiently integrate multiple algorithmic outputs in a unified way. We quantitatively analyze the retrieval accuracies of image sequences of many different algorithms. We also evaluate the ability of our algorithm to visualize the general users' reviews. We build a set of query text $\mathcal{Q}$ by selecting 100 story-like reviews that mainly describe the flow of the trip sequentially, from each of TRIPADVISOR and YELP datasets. Since the query reviews are text-only and have no groundtruth of image sequences, we employ crowdsourcing-based evaluation via Amazon Mechanical Turk (AMT).

Contributions. (1) To the best of our knowledge, this work is the first to address the ranking and retrieval of image sequences for long, multi-paragraph, natural language queries. We extend both input and output to more complex forms in relation to previous research: paragraphs instead of sentences and image sequences instead of individual images.

(2) We develop an image sequence retrieval method, built upon a structural ranking SVM with latent variables. In our experiments, we show that our method can flexibly incorporate different pieces of information about the text and image structure.

(3) We evaluate our method with large unstructured *Disneyland* dataset, consisting of 10K blog posts with 120K associated images, and 6K photo streams of 540K images. With quantitative evaluation and user studies via Amazon Mechanical Turk, we show our approach is practical in visualizing natural language text written by actual users.

[1] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *JAIR*, 47:853–899, 2013.

[2] Thorsten Joachims. Training Linear SVMs in Linear Time. In *KDD*, 2006.

[3] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *NIPS*, 2011.

[4] Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. Grounded Compositional Semantics for Finding and Describing Images with Sentences. In *TACL*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.