CISC: Clustered Image Search by Conceptualization

Kaiqi Zhao [#1], Enxun Wei [#1], Qingyu Sui [#1], Kenny Q. Zhu [#2] [*], Eric Lo [†3]

[#] *Shanghai Jiao Tong University*
[1] {kaiqi_zhao,nxun,sqybi}@sjtu.edu.cn
[2] kzhu@cs.sjtu.edu.cn
[†] *Hong Kong Polytechnic University*
[3] ericlo@comp.polyu.edu.hk

ABSTRACT

Clustering of images from search results can improve the user experience of image search. Most of the existing systems use both visual features and surrounding texts as signals for clustering while this paper demonstrates the use of an external knowledge base to make better sense out of the text signals in a prototype system called CISC. Once we understand the semantics of the text better, the result of the clustering is significantly improved. In addition to clustering the images by their semantic entities, our system can also conceptualize each image cluster into a set of concepts to represent the meaning of the cluster.

Categories and Subject Descriptors

H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval

General Terms

Algorithms

Keywords

Clustering, Image Search, Conceptualization

1. INTRODUCTION

Images are the most rich source of multimedia data on the Web. Present keyword-based search technology and information extraction techniques enable effective image search by keywords. For example, if one enters the keyword "kiwi" on Google Image search engine we got a whole screen of images which are associated with the term "kiwi", according to Google (see Figure 1). However, the term "kiwi" refers to at least two different entities: *kiwi the bird* and *kiwi the fruit*. Google apparently doesn't distinguish between these two entities and it mixes their images together. The ranking of these images is probably determined by the relevance ranking of the web pages that contain these images. Like Google, most other image search engine also return a mixed list of images without classifying the results by their different senses. Ambiguous entities like "kiwi" are abundant in real-world. For example, there are at least three different entities for "Polo", four for "Andrew Appel", six for "Anderson" and over ten for "Lei Zhang" on Google Image.

But when a user search for the term "kiwi", it is quite likely that he's looking for either the bird or the fruit but not both. Having to scroll through pages of images to locate the images that he wants is not a very good user experience. A better search result interface would be to divide the result window into sections, where each section contains only a number of images of the same type of entities.

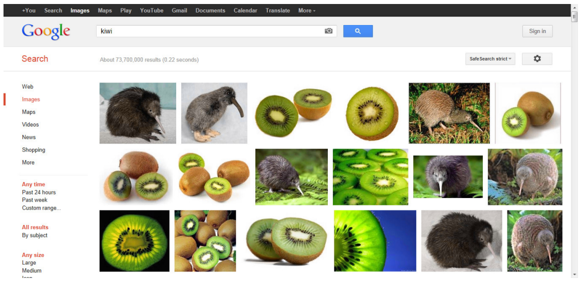


Figure 1: Search result of keyword "kiwi" on Google Image

User can then pick a section of interest and expand it further to see more images of that entity.

This paper is concerned with the problem of clustering a set of images indexed by the same keyword on an image search engine into multiple groups, each of which containing a distinct entity.

There are three existing approaches to image clustering. The first is "content-based" approach which uses only visual signals from the images themselves [6]. Most content-based image clustering approaches extract low-level visual features such as color and basic shapes from the images and use these to construct feature vector for similarity comparison. The problem with these approaches is that web images about the same entity can be very diverse and low level visual cues are often inadequate to capture the semantic commonality among the images. Take the images of the kiwi birds in Figure 1 for example, the background, color, pattern, brightness, contrast and orientation of the images can be quite different from image to image. An alteration of the above approach leverages visual object recognition and uses the object annotations for clustering. However, it cannot succeed in our problem because object recognition requires supervised learning and works only for a limited types of objects which have been manually labeled. Even for those manually trained object recognition tasks, the state-of-the-art techniques [7] still suffers from low accuracy.

The second is "context-based" approach which uses only the surrounding text, URL or tags of the images as feature space [2]. The challenge here is 1) finding the right context and 2) modeling the context. Existing techniques use visual cues from the web page to segment the page into blocks and uses the block that contains the image as its context. To model the context, all previous work uses variants of bag of words model which has limited capability of capturing the semantics of the text.

The third is a hybrid approach which combines visual features with textual features [5]. Because of the different nature in these two feature sources, there's no easy way of creating a combined

[*] Kenny Q. Zhu (corresponding author) is partially supported by NSFC Grant 61100050 and 61033002.

/ uniform signal for an image or using one single similarity function in the clustering. As such, existing methods often adopts a two-phase approach, taking advantage of the visual signals in one phrase, and the text signals in another. However, on many occasions, visual feature similarity and textual similarity of the context can conflict with each other so the above two-phrase approach can be counter-productive.

In this paper, we present a prototype system that takes the context-base clustering to the next level. We take this approach because 1) there are two possibly conflicting sources of signals for a given image: the visual signal and the text signal and there is no simple way to combine or reconcile them; and 2) text signals are more reliable and can be better captured by our novel techniques. In our framework, we extract two types of context of a given image: the HTML <IMG> tag context which consist of the URL and ALT name of the image, and the surrounding plain text context. We then conceptualize these context into a list of weighted Wikipedia concepts and cluster the images based on these concepts. We adopt a tri-stage clustering algorithm to obtain image clusters with high accuracy.

This paper hence makes three main contributions: 1) We developed a method to extract and conceptualize the image context using an external knowledge source; 2) Our novel tri-stage clustering algorithm yields clusters on benchmark web images with high accuracy; 3) The prototype system also characterises each cluster with a ranked list of concepts.

2. FRAMEWORK

In this section, we introduce a novel image clustering approach using text features. First, we extract relevant context of an image using a modified *sibling algorithm* [1]. Second, with those high quality context, we perform a *conceptualization* process on the context to gather its abstract semantics. Finally, we cluster images using a tri-stage clustering approach.

2.1 Context Extraction

The context extraction process can be described as a function that takes as input a web page, a image which is embedded in the page and a query string which produces the page and the image, and returns the context which includes two parts: the image tag context and the plain text context. Image tag context is essentially the URL and the ALT name (label) of the image in the <IMG> tag. The plain text context comes from two sources, one from the text surrounding the image and the other from the text surrounding the query string.

The text surrounding the image is very important because the most significant signals which can help us identify an image are often represented close to the image. We adopt a fast sibling based method to extract this context. First we select the parent of the image node in the HTML document. If this node has any readable text as its child then return its text as the context. Otherwise we select the parent of this node recursively until it has meaningful text or reach the root of the document. A window is also involved to restrict the length of context. This simple but reasonable method is shown to be effective in our experiments.

Context around the query string provides additional related semantic signals which may be far away from the image but close to the search term. The reason we employ this additional context is because the context surrounding the image is likely to be an accurate description of that image but not always enough to distinguish this entity from other entities of the same name.

Both image context and query string context are extracted in a bottom-up expanding procedure and the we require that every plain text fragment to contain at least one term which is a title of a Wikipedia article (a.k.a. Wikipedia concept). Furthermore, in the expansion procedure, we employ a heuristic that limits the context to a list item if we detect that the image or query string is located inside a list structure.

2.2 Conceptualization

Once we gather the necessary context of a web image, the next step is to convert the context into a list of *weighted concepts* in a process called *conceptualization*. We do this because concepts provides abstraction and therefore high level understanding of human language. Therefore concepts are better at capturing the semantics of the context. In this work, we use Wikipedia as an external knowledge source to provide the domain of all concepts. The current version of Wikipedia contains over 4 million articles, each of which describes a concept which is usually the title of the article. We conceptualize extracted context in two steps: *Wikification* and *Scoring*.

Wikification [3] is a process that links the noun phrases in a plain text to the corresponding Wikipedia articles. This is similar to word sense disambiguation in that it assigns a concept (sense) to a noun phrase. Because links can be sparse in Wikipedia data, we add links to as many unlinked terms as possible in Wikipedia using an iterative method.

Once the context is "wikified", it can be represented by the list of Wikipedia concepts. Next we compute a score called *CF-IDF* (concept frequency and concept inverse document frequency) to present the relative importance of each concept with respect to the context. *CF-IDF* is similar to the well-known *TF-IDF* except we compute the frequency of a concept in the Wikipedia corpus rather than the frequency of the surface forms (terms). To obtain the concept *IDF*, we scan all of the articles in Wikipedia corpus, and for each concept, we count the number of documents in which the concept appears as a link. *CF-IDF* is computed as follows:

$$\text{CF-IDF}(c, d) = cf(c, d) * log\frac{N}{df(c)} \tag{1}$$

where c is a Wikipedia concept, d is the given document and N is the total number of Wikipedia articles. $cf(c, d)$ stands for frequency of c in d, and $df(c)$ for the number of Wikipedia articles in which c occurs.

2.3 Tri-Stage Image Clustering

We propose a tri-stage image clustering method to process the conceptualized contexts. The three stages are tag context clustering, text context clustering and expansion clustering. Each stage form bigger clusters by merging the clusters formed in previous stage.

Tag context clustering: The first stage clusters images by the most reliable signals, i.e., the tag context because <IMG> tags directly describes the images in question. We analyze the URL to obtain Wikipedia concepts. Since the URL may contain random strings sometimes, we train a classifier to filter the random strings. Then we detect Wikipedia concepts from the URL and the image label to build a concept vector, in which each dimension is a Wikipedia concept with CF-IDF as its weight. Finally we cluster the vectors using Hierarchical Agglomerative Clustering(HAC) algorithm with cosine similarity.

Text context clustering: The tag context clustering forms many small and tight clusters. Next, we concatenate the text contexts of all images from a cluster to form the text contexts for clusters. This stage also employs HAC algorithm to further merge conceptually similar clusters.

Expansion clustering: The above two clustering steps are based on the exact concepts extracted from the web pages. To discover more hidden signals, we expand our contexts by taking the top-k ranked concepts (by CF-IDF score) of each cluster, and replace each of the top concepts by the linked concepts in their Wikipedia articles. The result will be an expanded concept vector for each cluster. We then further merge some of the clusters by the new concept vectors.

3. PRELIMINARY RESULTS

We present some preliminary results on web image clustering. In the following subsections, we first introduce the data set, then the evaluation metrics and finally compare the accuracy of our approach with the baseline bag-of-word context clustering as well as a state-of-the-art hybrid clustering method.

3.1 Data Set

We obtain a benchmark data set of web image data from *Google Image Search*, by querying 56 ambiguous entity names, such as *kiwi*, *pluto* and *explorer*. For each entity, we download the top 100 images along with the original web pages of the images. The whole data set contains 5,600 web pages/images in total.

3.2 Evaluation Metric

We adopt two metrics to evaluate the result of image clustering: F_1 and *NMI*. F_1 score combines *Purity* and *Inverse Purity* of the clusters. Similar to the F_1 score used in information retrieval task, F_1 score is computed in the following way:

$$F_1(C,L) = \frac{2 \times Purity(C,L) \times IPur(C,L)}{Purity(C,L) + IPur(C,L)} \qquad (2)$$

where $IPur$ stands for *Inverse Purity*. *Purity* measures the quality within a cluster. It is computed as follows:

$$purity(C,L) = \frac{1}{N}\sum_i \max_j |c_i \cap l_j|, c_i \in C \; and \; l_j \in L \qquad (3)$$

where C is the clusters and L is the set ground truth labels. *Inverse purity* is computed by inverting the ground truth and the resulting clusters in Equation 3, i.e. $purity(L,C)$.

NMI (Normalized Mutual Information) describes the amount of common information between the resulting clusters and the ground truth. Equation 4 shows the definition of NMI score, where Equation 5 stands for mutual information $I(C,L)$.

$$\text{NMI}(C,L) = \frac{I(C,L)}{(H(C)+H(L))/2} \qquad (4)$$

$$I(C,L) = \sum_i \sum_j \frac{|c_i \cap l_j|}{N} log \frac{N|c_i \cap l_j|}{|c_i||l_j|} \qquad (5)$$

$H(C)$ and $H(L)$ are entropy of clusters C and ground truth L respectively.

3.3 Image Clustering Accuracy

We manually label 14 entities from the data set for testing. Table 1 shows the comparison of the tri-stage clustering(TSC) algorithm against the hybrid clustering method MMCP [5] and a baseline bag-of-words(BOW) approach.

BOW has very poor F1 and NMI score, since only considering the concepts appears in the context(sometime the context is very short) without external knowledge will have insufficient signals to merge the small clusters. Compared to BOW, the TSC algorithm bring much more signals to identify the similarity between two clusters. As shown in the table, F1 and NMI score have around 0.3 and 0.18 improvement respectively.

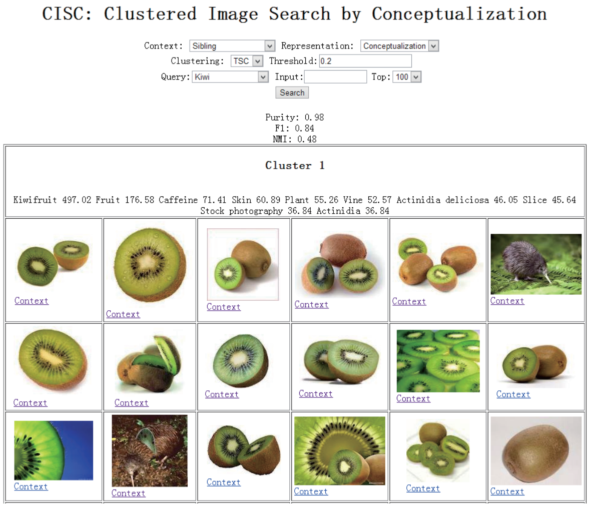
In our experiment, MMCP algorithm was implemented using two modalities: color histogram and textual feature (bag of words). As shown in Table 1, our system outperforms this state-of-the-art framework by significant margins. We thus argue that conceptualization of the image context is more powerful than combination of simple text and visual features.

Table 1: Results on different queries with different methods

	TSC		MMCP		BOW	
Query	**F1**	**NMI**	**F1**	**NMI**	**F1**	**NMI**
Amazon	**0.87**	**0.49**	0.46	0.42	0.3	0.23
Apple	0.72	0.29	**0.75**	**0.5**	0.3	0.25
Bean	**0.92**	**0.85**	0.47	0.51	0.61	0.58
Emirates	**0.82**	0.2	0.43	**0.42**	0.52	0.3
Explorer	**0.93**	**0.67**	0.62	0.42	0.76	0.47
Focus	**0.9**	**0.69**	0.53	0.43	0.43	0.32
Kiwi	**0.84**	**0.48**	0.53	0.42	0.53	0.34
Malibu	**0.89**	**0.69**	0.64	0.44	0.7	0.48
Pluto	**0.86**	**0.45**	0.61	0.31	0.61	0.25
Polo	**0.87**	**0.63**	0.56	0.48	0.74	0.5
Sante Fe	**0.87**	**0.55**	0.49	0.37	0.64	0.36
Tick	**0.78**	**0.4**	0.65	0.38	0.31	0.27
Time	0.59	0.32	0.59	**0.7**	0.41	0.22
Tucson	**0.94**	**0.71**	0.53	0.41	0.67	0.38
Avg.	**0.84**	**0.53**	0.56	0.44	0.54	0.35

4. DEMONSTRATION

Our demo system is implemented in .NET environment on a workstation with 3.2 GHz Dual-core i5 and 14GB RAM. A snapshot of our system is shown in Figure 2. Our system provides several variants of context extraction and clustering algorithms for comparison purposes. Table 2 shows the configurable options for users. If the user choose TSC as clustering method, the Representation option is fixed to Conceptualization, and the Context options is fixed to Sibling. Our system can provide 19 kinds of combination of methods for comparison.



CISC: Clustered Image Search by Conceptualization

Context: Sibling Representation: Conceptualization
Clustering: TSC Threshold: 0.2
Query: Kiwi Input: Top: 100

Purity: 0.98
F1: 0.84
NMI: 0.48

Cluster 1

Kiwifruit 497.02 Fruit 176.58 Caffeine 71.41 Skin 60.89 Plant 55.26 Vine 52.57 Actinidia deliciosa 46.05 Slice 45.64 Stock photography 36.84 Actinidia 36.84

Figure 2: System snapshot

For the context extraction process, we provide options such as *Whole*, *Sibling*, and *Sibling(Image only)*. The *Whole* option uses the text of the whole page as context. *Sibling* consider both of the

Table 2: Various configurations of the demo system

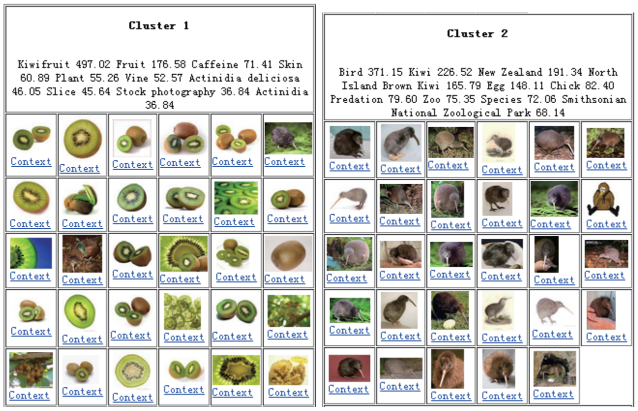
Context	Representation	Clustering
Whole	BOW	AP
Sibling	Conceptualization	HAC
Sibling(Image only)	BOC	TSC

image context and query string context, which is proposed in our system. *Sibling (Image only)* provides an option of only using the text around the image as context.

Similar to context extraction, there are three context representations *BOW, Conceptualization, BOC* in the system. BOC stands for "bag-of-concepts", in which we detect Wikipedia terms without explicitly disambiguate the sense of the terms. User can also select the clustering method such as TSC, *Affinity Propagation(AP)*[4] and HAC, and tune the threshold for each clustering method.

We offer 56 query strings for testing. These queries were downloaded in advance since the Google Image search engine is not always available in our part of the world. Besides the 56 pre-loaded queries, user can enter new queries in the search box. But the download process may take a long time. The *Top* option indicates that we are processing the top K images retrieved from search engine.

Figure 3 shows two largest clusters of "kiwi" from a total of 100 images. The accuracy (purity, F1, NMI) of the result are shown on the top, and the clusters are listed below. In this particular result, only one kiwi bird image is incorrectly grouped into the kiwifruit cluster since its text is more about kiwifruit than about kiwi bird.



Cluster 1

Kiwifruit 497.02 Fruit 176.58 Caffeine 71.41 Skin 60.89 Plant 55.26 Vine 52.57 Actinidia deliciosa 46.05 Slice 45.64 Stock photography 36.84 Actinidia 36.84

Cluster 2

Bird 371.15 Kiwi 226.52 New Zealand 191.34 North Island Brown Kiwi 165.79 Egg 148.11 Chick 82.40 Predation 79.60 Zoo 75.35 Species 72.06 Smithsonian National Zoological Park 68.14

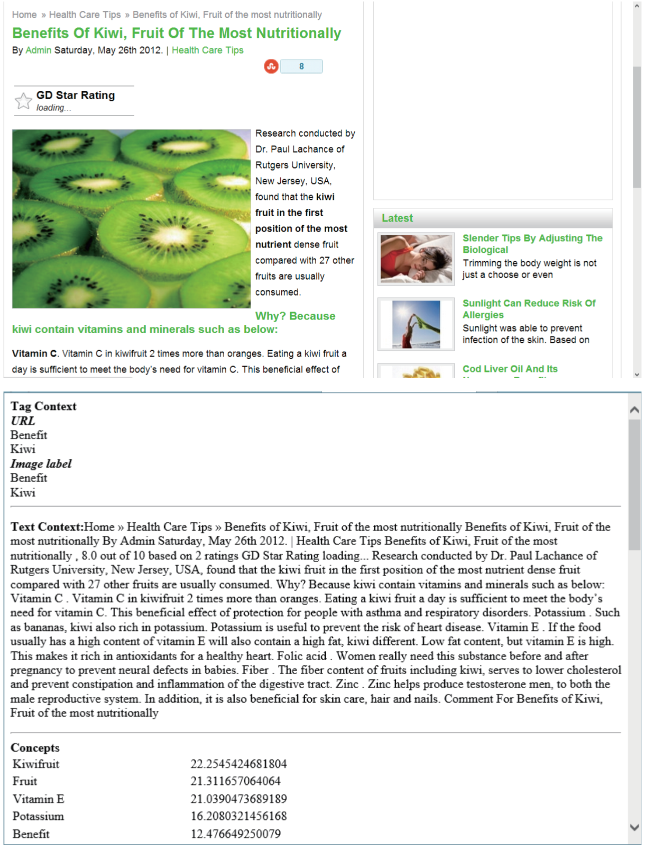
Figure 3: Two clusters of kiwi

In our system, for each cluster, we list the top 10 most representative concepts to describe each cluster. For example, the left sub-figure of Figure 3 shows a cluster about *Kiwifruit, Fruit*, etc. while the right sub-figure shows a cluster about *Bird, Kiwi*, etc. (Notice that "Kiwi" is the term for kiwi the bird in Wikipedia.)

The demo system also keeps track of the context extraction process. When the user clicks on an image in the result set, the system will open the original web page in the browser. If the user clicks on the *context* link, the system will show the contexts (both tag and text context). At the end of the context page, there is a ranked list of Wikipedia concepts extracted from the context. Figure 4 shows the original web page (top) and the context (bottom) of an image of kiwifruit.

5. REFERENCES

[1] S. Alcic and S. Conrad. Measuring performance of web image context extraction. In *MDMKDD*, pages 8:1–8:8, 2010.



Home » Health Care Tips » Benefits of Kiwi, Fruit of the most nutritionally

Benefits Of Kiwi, Fruit Of The Most Nutritionally

By Admin Saturday, May 26th 2012. | Health Care Tips

8

GD Star Rating
loading...

Research conducted by Dr. Paul Lachance of Rutgers University, New Jersey, USA, found that the **kiwi fruit in the first position of the most nutrient** dense fruit compared with 27 other fruits are usually consumed.

Why? Because kiwi contain vitamins and minerals such as below:

Vitamin C. Vitamin C in kiwifruit 2 times more than oranges. Eating a kiwi fruit a day is sufficient to meet the body's need for vitamin C. This beneficial effect of

Latest

Slender Tips By Adjusting The Biological
Trimming the body weight is not just a choose or even

Sunlight Can Reduce Risk Of Allergies
Sunlight was able to prevent infection of the skin. Based on

Cod Liver Oil And Its

Tag Context
URL
Benefit
Kiwi
Image label
Benefit
Kiwi

Text Context: Home » Health Care Tips » Benefits of Kiwi, Fruit of the most nutritionally Benefits of Kiwi, Fruit of the most nutritionally By Admin Saturday, May 26th 2012. | Health Care Tips Benefits of Kiwi, Fruit of the most nutritionally , 8.0 out of 10 based on 2 ratings GD Star Rating loading... Research conducted by Dr. Paul Lachance of Rutgers University, New Jersey, USA, found that the kiwi fruit in the first position of the most nutrient dense fruit compared with 27 other fruits are usually consumed. Why? Because kiwi contain vitamins and minerals such as below: Vitamin C . Vitamin C in kiwifruit 2 times more than oranges. Eating a kiwi fruit a day is sufficient to meet the body's need for vitamin C. This beneficial effect of protection for people with asthma and respiratory disorders. Potassium . Such as bananas, kiwi also rich in potassium. Potassium is useful to prevent the risk of heart disease. Vitamin E . If the food usually has a high content of vitamin E will also contain a high fat, kiwi different. Low fat content, but vitamin E is high. This makes it rich in antioxidants for a healthy heart. Folic acid . Women really need this substance before and after pregnancy to prevent neural defects in babies. Fiber . The fiber content of fruits including kiwi, serves to lower cholesterol and prevent constipation and inflammation of the digestive tract. Zinc . Zinc helps produce testosterone men, to both the male reproductive system. In addition, it is also beneficial for skin care, hair and nails. Comment For Benefits of Kiwi, Fruit of the most nutritionally

Concepts

Kiwifruit	22.2545424681804
Fruit	21.311657064064
Vitamin E	21.0390473689189
Potassium	16.2080321456168
Benefit	12.476649250079

Figure 4: Context extraction

[2] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen. Hierarchical clustering of www image search results using visual, textual and link information. In *MM*, pages 952–959, 2004.

[3] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of EMNLP-CoNLL*, volume 2007, pages 708–716, 2007.

[4] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:2007, 2007.

[5] Z. Fu, H. H.-S. Ip, H. Lu, and Z. Lu. Multi-modal constraint propagation for heterogeneous image clustering. In *MM*, pages 143–152, 2011.

[6] S. hua Zhong, Y. Liu, and Y. Liu. Bilinear deep learning for image classification. In *MM*, pages 343–352, 2011.

[7] L.-J. Li, R. Socher, and F.-F. Li. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, pages 2036–2043, 2009.