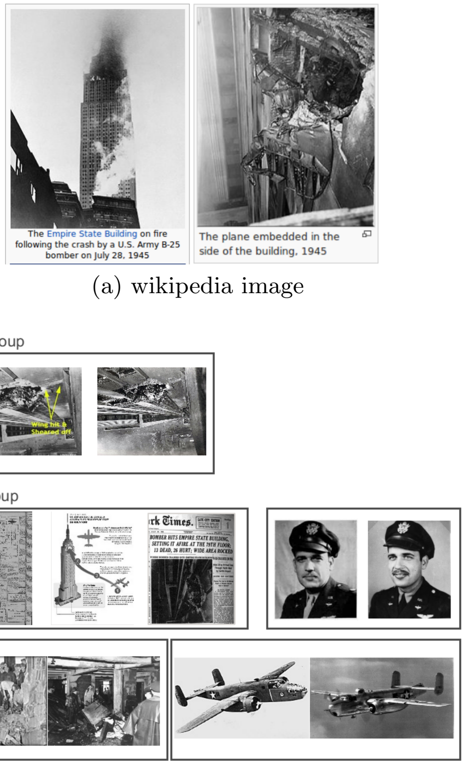
Image Profiling for History Events on the Fly

Jia Chen* [1], Qin Jin† [2], Yong Yu[1], Alexander G. Hauptmann[3]

[1]Dept. of Computer Science & Engineering, Shanghai Jiao Tong University
{chenjia, yyu}@apex.sjtu.edu.cn
[2] School of information, Renmin University of China
qjin@ruc.edu.cn
[3]School of Computer Science, Carneige Mellon University
alex@cs.cmu.edu

ABSTRACT

History event related knowledge is precious and imagery is a powerful medium that records diverse information about the event. In this paper, we propose to automatically construct an image profile given a one sentence description of the historic event which contains where, when, who and what elements. Such a simple input requirement makes our solution easy to scale up and support a wide range of culture preservation and curation related applications ranging from wikipedia enrichment to history education. However, history relevant information on the web is available as "wild and dirty" data, which is quite different from clean, manually curated and structured information sources. There are two major challenges to build our proposed image profiles: 1) unconstrained image genre diversity. We categorize images into genres of documents/maps, paintings or photos. Image genre classification involves a full-spectrum of features from low-level color to high-level semantic concepts. 2) image content diversity. It can include faces, objects and scenes. Furthermore, even within the same event, the views and subjects of images are diverse and correspond to different facets of the event. To solve this challenge, we group images at two levels of granularity: iconic image grouping and facet image grouping. These require different types of features and analysis from near exact matching to soft semantic similarity. We develop a full-range feature analysis module which is composed of several levels, each suitable for different types of image analysis tasks. The wide range of features are based on both classical hand-crafted features and different layers of a convolutional neural network. We compare and study the performance of the different levels in the full-range features and show their effectiveness on handling such a wild, unconstrained dataset.


The Empire State Building on fire following the crash by a U.S. Army B-25 bomber on July 28, 1945
The plane embedded in the side of the building, 1945
iconic group
Wing penetration
Wing hit & Sheared off
facet group

(a) wikipedia image

(b) generated image profile (images are grouped by boxes)

Figure 1: Illustration of image profiling for history event on the fly: input text sentence is "At 9:40 a.m. on Saturday, July 28, 1945, a B-25 Mitchell bomber, piloted in thick fog by Lieutenant Colonel William Franklin Smith, Jr., crashed into the north side of the Empire State Building, between the 79th and 80th floors.".

General Terms

H.3.3 Information Search and Retrieval

Keywords

image profiling; history event

1. INTRODUCTION

"What is history? An echo of the past in the future; a reflex from the future on the past" once said by Victor

*The work was done when the author was visiting Renmin University of China.

†Qin Jin is the corresponding author

MM'15, October 26–30, 2015, Brisbane, Australia.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806242.

Hugo [2]. Knowledge of history is valuable. Some of this knowledge is recorded in text form containing exact information about“where”, “when”, “who” and “what” (the 4Ws of an event) while the majority of information is recorded in image media such as photo snapshots, art paintings and maps. There currently are many well organized text resources about historic events. However, images embody a wealth of information recorded and transmitted by various means including paper, artistic expression (e.g. painting) and camera. We find there are relatively few organized image resources about historic events but there are many image resources related to these events scattered around the web which can be found through a combination of different keyword queries on image search engines. This big gap between organized image resources for historic events and available image resources online makes automatic image profiling of historic events desirable and valuable. An example is the famous “B-25 Empire State Building crash” event. On Wikipedia, there is an individual page describing this event [1] with two illustrating images as shown in Figure 1(a). The image profile constructed by our algorithm from online resources is shown in Figure 1(b). It not only covers far more photos with acceptable quality, but also different genres of images such as newspaper reports and diagrams. Such historic event profiles facilitate better understanding and preservation of historical knowledge.

In our proposed image profiles, photos function as vivid snapshots of historic events, paintings provide an artistic subjective view, while document/maps supplement background information. These profiles benefit a wide range of cultural preservation and curation related applications covering history education, wikipedia enrichment, material supplement for museums, and false historic event photo detection.

The input requirement of the image profiling system is extremely simple: a sentence describing a historic event containing the 4Ws. This simple requirement makes it extremely easy to apply our approach to many types of historic events and the resulting image profile corpus can grow rapidly. However, there are two major challenges in building an image profile for an event.

Unconstrained Image Genres: Unlike most computer vision datasets which focus only on photos, photos are just one of many image genres that record historic event related materials. Related material can be in the form of images of a scanned document, an original manuscript, a hand crafted map, an oil painting, a portrait, a sketch and so on. We divide images into three general genres: documents/maps (recording history on paper), paintings (recording history through artistic expression) and photos (recording history with cameras). Furthermore, within each genre, there are many variants. The document/map genre includes scanned documents, original manuscripts and hand-drawn maps. Paintings can include a variety of styles from Renaissance to Realism. The variations in photo genres have already been well recognized [4]. Because of such diversity, genre classification requires a full range of feature analysis methods from low-level color distribution and edge detection to high-level semantic content identification. Our approach utilizes the multiple layers of a neural network image processing model [14][9] to effectively classify genres.

Diverse Image Content: Most well-known computer vision research datasets focus on a single aspect of computer vision such as faces[11], objects[5], scenes[35] and so on. However, in historic event image profiling, the content of related images are unconstrained. Furthermore, even for the same event, the related images can be quite diverse, describing different facets of the event. One goal of our system is to aggregate images portraying various facets of an event to provide diversity in the profile. We solve this challenge by grouping images at two levels of granularity: iconic image grouping and facet image grouping. We achieve these two-level groupings via ***iconic image detection*** and ***image diversification***, respectively. Iconic images are classic event images that have been propagated to multiple places on the web. In the propagation process, images may be cropped or edited, which makes iconic image detection even more challenging. Image diversification for historic events is more difficult than image search result diversification of object queries since the visual appearance of images within the same group is more varied and images only share similar semantics of content. The two levels of granularity require different types of feature analysis ranging from near exact matching to content semantic similarity.

The solution for both challenges requires a broad-range of feature analysis from low-level pixels to the high-level content semantics. Our approach leverages the hierarchy of layers in a convolutional neural network where different layers can be considered as different levels of feature analysis. Technically, we view each layer’s output in a convolutional neural network as a feature map extracted on a dense grid. We construct different level feature descriptors upon different layer’s raw features and fuse these feature descriptors to solve the above two challenges. We also include classical hand-crafted features as the bottom level in our full-range feature approach. We do not apply fine-tuning to adapt the convolutional neural network in this work for two reasons: First, fine-tuning needs a significant amount of labelled image data to work well. This is not available in our image profiling setting, since the amount of available images is relatively small. Second, our feature extractor view of a convolutional neural network might only need pre-trained models. Different from work in [6][24] which focuses on directly using the output of the last two hidden fully connected layers for high-level vision tasks, we focus on constructing a wide range of features to solve different types of visual classification tasks such as genre classification and image grouping.

We make the following contributions:

1. we propose a novel system to automatically construct an image profile given a one-sentence description of the historic event, which has many potential applications related to cultural preservation and curation. For this, we develop a flexible approach making efficient use of DCNNs to facilitate classification and labeling for this work.
2. we construct a full-range feature analysis framework to study the best level or level combinations for different tasks such as shown in Figure 3. This helps us to better understand differences between tasks and guides us to improve the performance accordingly.
3. we apply this framework to genre classification and find that level 5 (the middle level) feature works best, which is very different from traditional object detection tasks where high level features usually work best.
4. we also apply this framework for image grouping and find that the level 0 (hand-crafted) feature works best for iconic

image grouping while level 6 (high level) works best for facet image grouping.

The remainder of the paper is organized as follows. Section 2 presents related work. Section 3 presents the detailed process of image profiling for historic events based on the full range of features. Section 4 discusses detailed evaluation on image profiling and does case study to show the effectiveness of the proposed full-range of features. Section 5 concludes the paper.

2. RELATED WORK

2.1 Genre Classification

Genre classification has not been extensively studied. The most related work is distinguishing paintings from photographs [4]. A bunch of hand-crafted features, most of which are based on color, are designed in work [4]. The features are shown to be effective on a dataset composed of paintings from archive websites and photos from photography library website. However, in our collected historic events dataset, the designed features don't work well. Based on our analysis, there are two reasons for the dramatic performance drop using these features on our dataset. First, the images in our dataset are wild and much more difficult, i.e. there are few typical paintings and few typical photos. Color based features do not have much discriminative power for such difficult dataset. Second, these features seem to be over-tuned for the dataset used in work [4]. We propose to design a wide range of features based on the rich information contained in the different layers of a pre-trained convolutional neural network to solve the genre classification problem. We show that such features are robust for a wild genre classification dataset.

There are much work on processing document/map images [29][13] but little on distinguishing it from other genres. In historic event related images, document images include not only cleanly scanned black-white document but also photos of old faded newspapers, hand written manuscripts and hand-crafted colored maps. Given the diversity of images in document genre, it is not a trivial task. We apply our wide range of features to classify document/map, painting and photo genres simultaneously.

2.2 Image Search Result Diversification

Image search result diversification is useful to improve the quality of text-based image retrieval [32][30]. Most of the queries studied in these papers are object queries with ambiguity such as apple and jaguar. Image results for such queries can be diversified based on low-level features such as color and texture. However, diversified image grouping for historic events relies much more on the content semantic and images within the same group may be diverse on low-level features. We show that low-level color, edge and texture features don't work well in this challenging task. We propose to exploit the semantic expressive power of the wide range of features combined with region proposal to match images within a group.

2.3 Convolutional Neural Network

Convolutional neural network has been shown to be the state-of-the-art on a variety of high-level computer vision tasks such as object classification [14] and object detection [9]. Pre-trained convolutional neural network has also been used as a feature extractor for generic visual recognition [6][24]. However, the focus is on the semantic discriminative power via directly using the output of last several layers for recognition tasks. Features based on rich information of other layers for different vision tasks have not been well unexplored. Work in [6] only considers the output of last three layers as the authors think that only the output of these three layers are activations fully propagated through the convolutionl layers of the network, which is true only in direct use of layer outputs. Work in [24] evaluates all the layer outputs only in the recognition task and then only applies the output of last two hidden layers in other vision tasks. Different from these works, we design a wide range of features based on the output of different layers in the network rather than directly using the output layer. They are more compact particularly for intermediate convolutional layer outputs. Based on this compact representation, we compare and study the performance of different levels in the wide range of features on genre classification and diversified image grouping on our challenging historic event dataset.

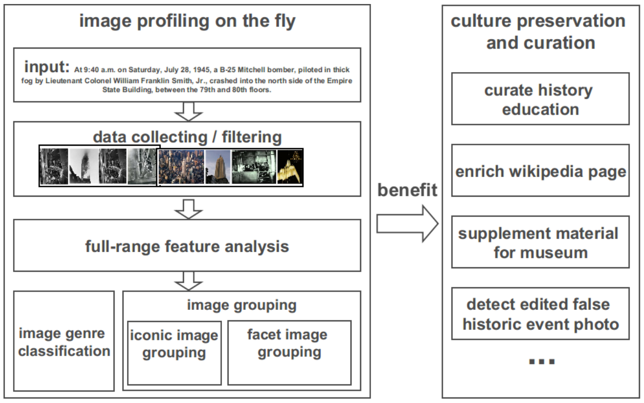


image profiling on the fly
input: At 9:40 a.m. on Saturday, July 28, 1945, a B-25 Mitchell bomber, piloted in thick fog by Lieutenant Colonel William Franklin Smith, Jr., crashed into the north side of the Empire State Building, between the 79th and 80th floors.
data collecting / filtering
full-range feature analysis
image genre classification
image grouping
iconic image grouping
facet image grouping
benefit
culture preservation and curation
curate history education
enrich wikipedia page
supplement material for museum
detect edited false historic event photo
...

Figure 2: Framework of image profiling

3. SYSTEM DESCRIPTION

The overall system framework is composed of four modules: data collection and filtering, full-range feature analysis, image genre classification and image grouping as shown in Figure 2. In this section, we present the detailed implementation for each module in our image profiling system. We first describe the data collection and filtering procedure, which gathers data from the web and filters out clearly irrelevant images. We then explain the full range of feature analysis, which will be used in both genre classification and image grouping modules. Later in the genre classification, we show the robustness of the full range of features compared to hand-crafted features and study the performance of different levels in the wide range of features. Finally in the image grouping, we consider grouping at two levels of granularity: iconic image grouping and facet image grouping. Different versions of the iconic images will be categorized into a single group. Images that share similar content semantics are grouped into the same facet group and different facet groups correspond to different facets of a historic event (see also Figure 1).

3.1 Data Collection and Filtering

Data Collection: For each input sentence, we run named entity detection using Stanford NLP toolbox [20] and iden-

tify “who”, “where” and “when” elements. We also extract noun phrases, verbs and other named entities as key phrase candidates for query construction. Some examples of the key phrase candidates are highlighted in Table 1.
For each event, we construct queries by combining each keyword with the “when” and “where” elements to collect related images for the historic event.

Data Filtering: Even for a human being without very good background knowledge, it may be difficult to tell whether an image is really relevant to a specific event or not. Therefore, we only attempt to filter out clearly irrelevant images, which is a much easier task that does not require too much knowledge of the event. Based on observations of our collected data, a majority of definitely irrelevant images are contemporary landmark photos retrieved by the “where” element in the query. Thus we construct additional background queries which contain only the “where” element. Images retrieved by “where” only queries are classic contemporary landmark photos. We measure the similarity between a candidate image and these classic contemporary landmark photos using the level 6 in our full-range features. We remove images that are too similar to classic contemporary landmark photos according to a similarity threshold. We note that the data filtering can not be reduced to a simple color filtering task, which would initially seem to be a tempting thing to do. Since the images we are handling include different genres such as documents/maps , paintings and photos, we cannot simply remove colourful images even for events that happened before the invention of color camera. Furthermore, the definition of the relevance of an image to an event is vague and subjectively changes from person to person and event to event. Thus, we choose to remove only those contemporary images which are obviously irrelevant.

3.2 Full-range Feature Analysis

The full-range feature is composed of several levels. Level 0 is composed of all the hand-crafted features. Some are classic general purpose features such as sift [18]. Others are features designed for special purposes such as image quality. Features in level 0 will be introduced later when used.

Features in the rest levels are based on the output of different layers in a convolutional neural network. All the three state-of-the art convolutional neural networks (CNN) [14][36][28] share a similar layer architecture of 5 convolutional layers (conv1-5) and 3 fully-connected layers (full6-8). We do not consider the output of the last fully-connected layer in designing our full-range feature since it is the softmax for category prediction. We do not consider the output of the first layer either since it has been shown to be similar to edge features, which are covered by our features in level 0. For the remaining layers, we construct one level-feature from the output of each layer. The construction process for the level-features is different for the convolutional layer and fully-connected layer.

For a convolutional layer, its output is actually n channels of $m \times m$ feature maps. Each channel corresponds to a filter. Each element in the feature map corresponds to the response of that filter on a particular region in the original image. Directly using the output leads to a feature with very high dimensions ($n \times m \times m$), which is usually at the scale of hundreds of thousands. We construct a compact feature representation (21 dimension) by a three-level pyramid average pooling on grids 1x1, 2x2, 4x4 on each feature map.

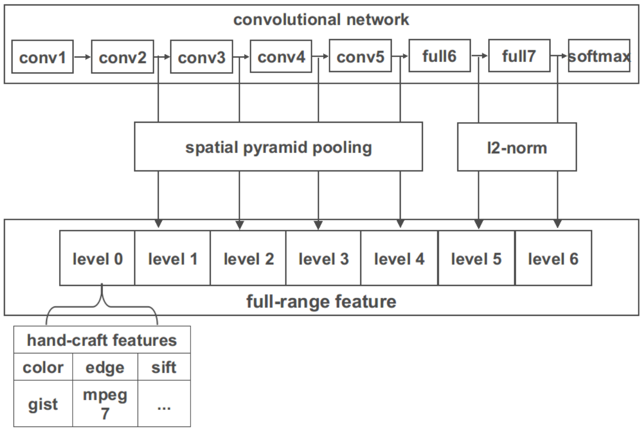


Figure 3: Construction of full-range features

Then we concatenate the compact feature representation by channels into a $21 \times n$ dimensional level-feature. For the two hidden fully-connected layers, we normalize their output by $l2-norm$ to get the corresponding level-features. In a summary, our full-range feature is composed of 7 levels. Level 0 is composed of all the hand-crafted features. The next 4 levels are constructed on the outputs of conv2-4 layers and the last 2 levels are constructed on the outputs of full6-7 layers. The construction of the full-range feature is summarized in Figure 3.

The dimension of each level-feature in the full-range feature is around tens of thousands, which is rather compact compared to that of commonly used HoG features. This enables us to enumerate the performance of different levels to find out best level feature for different tasks at a relatively low cost. In implementation, we construct the full-range feature based on the pre-trained CNN-M [33]. The full-range features can also be constructed upon other pre-trained CNNs as long as they follow the same architecture.

Level analysis of full-range features: In previous research of art, history and culture [25][17][38], authors typically design various hand-crafted features based on different principals and prior knowledge of certain domain. Different from these approaches, in our task the relevant information is available as a “wild and dirty” dataset, which makes hand-crafting specific features based on principles and prior knowledge almost impossible. Therefore, we build our solution as a mixture of hand-crafted general features and different layer outputs from a CNN. Lower levels usually capture pixel information while higher levels usually reflect soft content semantics. We apply level analysis by comparing performance of different level-features on the two challenges (genre classification and image grouping) in image profiling for historic events.

3.3 Image Genre Classification

We consider three genres: document/map, painting and photo in this work. Document/map genre includes scanned document photos, maps drawn by hand and other documentary records saved in image format. Painting genre includes oil painting, watercolour, sketch and other artistic records. Photo genre includes natural photos. To distinguish document/map from photo, we usually rely on the content semantics of the image. To distinguish paintings from photos, we rely on not only the content semantics but also low-level

Event ID	Where	When	Who	What (event description)
E12	Paris	1987	A.J. Hackett	1987: {*A.J. Hackett*} made {*one of his first bungee jumps*} from the top of the Eiffel Tower, using {*a special cord*} he had helped develop.
E66	Paris	1885	Victor Hugo	Prior to {*burial*} in {*the Pantheon*}, {*the body of Victor Hugo*} was exposed under {*the Arc*} during the night of 22 May 1885.

Table 1: Two historic event descriptions: key phrases used to construct queries are emphasized in italics

name	function
color edge	the removal of color eliminates more visual information from a painting than from a photograph of a real scene.
spatial variation of color	color changes to a larger extent from pixel to pixel in paintings than in photographs.
number of unique colors	paintings appear to have a larger color palette than photographs.
pixel saturation	paintings tend to contain a larger percentage of highly saturated pixels where as photographs contains more unsaturated pixels.
pixel distribution in RGBXY space	Paintings use a larger color palette and also have larger spatial variation of color, resulting in larger singular values for the covariance matrix of RGBXY space.

Table 2: List of level 0 features used in image genre classification

name	granualarity	function
sift	iconic image grouping	it is shown to provide robust matching across a a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination.
gist	facet image grouping	it is found to be effective in grouping images by perceptual similarity [10] and it describes global image structure.
mpeg7	facet image grouping	it has been widely used in image search result diversification [32] and it includes color layout, scalable color and edge histogram.

Table 3: List of level 0 features used in image grouping

features such as color edge and color distribution as stated in work [4].

The genre classification problem is a multi-class classification task, which is usually solved by ensembles of several one-vs-one or one-vs-all binary classifiers. In this work, we use one-vs-one binary classifier ensembles so that we can study which is the best level-feature to distinguish each pair of genres. Since the class number is 3, ensembles of one-vs-one binary classifiers is quite efficient. To be specific, we study feature levels from 0 to 6. In level 0, we implement the features described in work [4], which were used to distinguish paintings from photos. The list of implemented features is shown in Table 2. Thus, we only evaluate its performance in the painting vs photo classifier.

To create ensembles of binary classifiers, we apply three strategies as follows.

single level-feature: all the three binary classifiers use the same level-features and we select the best level-feature for each of them on the validation set.

late fusion: each binary classifier is tuned to its best level-features on the validation set and classification is achieved with each binary classifier using its own best level-feature.

early fusion: the best level-features for each binary classifier are selected on the validation set and fused into one feature at the feature construction stage.

3.4 Image Grouping

In image grouping, we organize images at two levels of granularity: iconic image grouping and facet image grouping. Iconic image grouping requires partial duplication within a group and is useful to find classical event images that have been propagated at multiple places on the web. Facet image grouping requires content semantics consistency within a group and is useful to construct different facets for the image profile of a historic event. Similar to genre classification, we also exploit different levels in the full-range features for grouping. Low-level features are expected to work well in iconic image grouping while facet image grouping needs high-level semantic features to get a descent performance.

The level 0 features used in image grouping involves SIFT [18], GIST [22] and Mpeg7 features [27].

3.4.1 Iconic Image Grouping

An iconic image is an image that is propagated to multiple places on the web. During the propagation, its content may be edited or cropped, but the major content is preserved. In historic event related images, iconic images appear frequently as famous classic photos for a historic event. Different from classic partial duplication solutions which usually focus on the dataset scale, the dataset scale of our partial duplication grouping is small and we focus on a high precision and recall since historic event photos are rare and precious. The dataset scale of iconic image grouping is the number of candidate images retrieved from several queries, which is typically around a few hundreds.

For all the level-features used in iconic image grouping, we follow a common pipeline as in [16]: matching image pairs and extracting connected components of matched pairs. The matching algorithms differ for different level-features.

level 0 *(SIFT + RANSAC):* we match points in an image pair based on descriptors and then estimate the underlying affine transformation using RANSAC [3] algorithm. In our task, we simplify the underlying affine transformation to 4 degree of freedom: scale change on x, y axis and translation on x, y axis. Rotation rarely happens in iconic image grouping in our collected dataset while cropping and editing is the major issue. Thus, such simplification helps reduce the hypothesis space and improve the estimated affine transformation quality. We get matched region from inlier pairs after affine transformation and its confidence can be measured by the number of matched pairs. Applying threshold on the inlier number can remove unmatched pairs.

level 0 *(SIFT + RANSAC + salient region):* This technique has proven to work well if the query region is given [23]. However, in iconic image grouping, the query region is not given beforehand and the matched region may be a background region as illustrated in Figure 4. This may be considered as a correct case in partial duplication grouping but incorrect for iconic image grouping. To solve this problem, we extract salient regions [37] and calculate the coverage of the matched region in the salient regions. Now the confidence of the iconic image pair is related to both the coverage and the inlier number. We also set a threshold for

genre	training	validation	test
doc/map	imagenet $(1,402)$	history (176)	history (404)
painting	archive (966)	history (259)	history (420)
photo	history $(1,000)$	history $(2,187)$	history $(3,427)$

Table 4: Split strategy of training, validation and test data for genre classification: the number in the bracket is the number of images from the corresponding source

level	doc/map vs painting	doc/map vs photo	painting vs photo
0	NA	NA	0.0/81.2/81.2
1	81.3/88.9/85.1	54.9/98.4/91.9	42.3/93.6/86.9
2	85.2/89.5/87.5	63.9/98.5/94.0	**50.1/94.1/88.9**
3	86.1/89.0/87.6	64.3/98.4/94.0	**50.0/94.6/88.9**
4	85.0/88.9/87.1	**65.7/98.5/94.3**	41.2/94.2/86.3
5	**88.1/87.8/87.9**	65.1/98.5/94.2	37.4/94.5/84.4
6	87.3/82.8/84.7	55.6/98.2/92.0	35.1/93.1/84.5

Table 5: Study of best level-features for genre classification on one-vs-one binary classifiers. The three numbers in each table element are precision of the first class, precision of the second class and overall precision.

coverage to further remove background matched pairs.
levels $1-4$ *(SPM):* the features in level $1-4$ are constructed by spatial pyramid pooling. Naturally, we can match them by spatial pyramid matching [15].
levels $5-6$ *(Cosine):* feature vectors are matched by calculating cosine similarity.

3.4.2 *Facet Image Grouping*

The related images of a historic event are diverse and different images are associated with different keyword facets of a historic event description. It is difficult for the users to find out what these images are about if they are mixed together. Thus it is necessary to group images into different groups for different event facets. This task is similar to image search diversification, in which image candidates of ambiguous queries are grouped to different facets. Such grouping usually uses different kinds of low-level features to group images and get reasonable results [32]. The queries studied are usually ambiguous about objects, which leads to the fact that resulting images containing a single major object and having coherent visual appearance within a group and distinguished visual appearance between groups. However, in historic event image profiling, images usually contain multiple objects and images within the same facet are more coherent on content semantics rather than their visual appearance. We use affinity propagation [8] for facet image grouping since it can handle different types of distances through pairwise similarities. We study the following distances on different level-features.
level 0 *(Mpeg7 + dist):* the similarity is the negative of the distance function implemented in LIRE [19].
level 0 *(GIST + euclidean):* the similarity is the negative of euclidean distance.
levels $1-4$ *(SPM):* the similarity is calculated using spatial pyramid matching [15].
levels $5-6$ *(Cosine):* cosine similarity between feature vectors is computed.

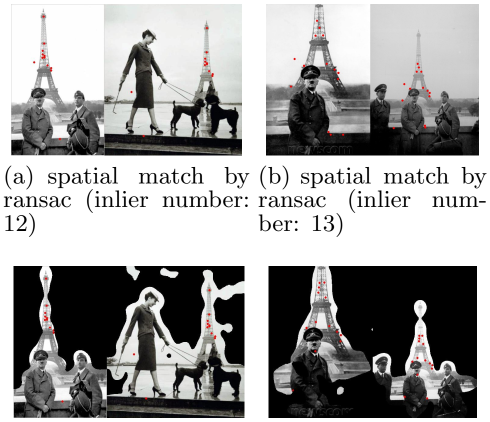

(a) spatial match by ransac (inlier number: 12)
(b) spatial match by ransac (inlier number: 13)
(c) salient region postprocess (coverage: 0.06)
(d) salient region postprocess (coverage: 0.21)

Figure 4: An example of adding salient region post-process for matching level 0 feature sift.

4. EXPERIMENT

4.1 Dataset

We parse events from Wikipedia pages[1]. The detailed process includes two steps: first, we get "where" by parsing pages of landmarks from the landmark list[2]; second, we get "when" by extracting sentences with year information in history sections on Wilipedia. Each sentence corresponds to the text description of one historic event. We discard events that happened after 2004 according to our definition of a historic event. We also discard events that happened too early, i.e. before the birth of photography(1816)[3]. In the end, 615 events and their corresponding descriptions are extracted. We crawl $40,912$ candidate images for these 615 events and $17,504$ images are classified as relevant images automatically.

4.2 Genre Classification

We randomly sample $6,693$ images from the automatically filtered history corpus and labelled genre ground-truth using Amazon Mechanical Turk (mturk)[4], which includes 580 doc/maps, 679 paintings and $5,434$ photos. We create the training data for document/map and painting genres by collecting additional images from other sources such as the imagenet and art archive websites. Specifically, $1,402$ images under category document (synset n06470073) and map (synset n03720163) from imagenet, 966 painting images from an art archive website (archive)[5] are collected in addition. The training and test data split strategy is listed in Table 4. Both the validation and test data are only composed of images from our historic event corpus.

We can see from Table 5 that the best feature lies in level 5, 4 and 3 for doc/map vs painting, doc/map vs photo and

[1] http://en.wikipedia.org/
[2] http://www1.i2r.a-star.edu.sg/~yzheng/landmark/1000_landmarks.html
[3] http://en.wikipedia.org/wiki/History_of_the_camera
[4] https://www.mturk.com/mturk/
[5] www.archive.com

method	precision
single level-feature (level 3)	65.3/47.4/93.5/85.2
late fusion	**70.6/48.1/92.9/86.0**
early fusion	**69.6/47.8/93.0/86.0**

Table 6: Evaluation on genre classification: The four numbers in each table element are precision of doc/map, precision of painting, precision of photo and overall precision. Late fusion is done by combining classifier doc/map vs painting on level 5, classifier doc/map vs. photo on level 5, classifier painting vs. photo on level 3. Early fusion is done by fusing level 3 and level 5 features

level	matching strategy	P_{pair}	R_{pair}
0	SIFT + RANSAC	0.95 ± 0.10	0.98 ± 0.02
0	SIFT + RANSAC + salient region	$\mathbf{0.98 \pm 0.06}$	$\mathbf{0.98 \pm 0.02}$
$1-4$	SPM	0.70 ± 0.15	0.83 ± 0.13
$6-7$	Cosine	0.33 ± 0.30	0.49 ± 0.29

Table 7: Evaluation on iconic image grouping

painting vs photo classifiers respectively, which is shallower than the best feature level widely reported in recognition tasks [28]. Furthermore, the best feature layer for painting vs photo classifier is shallower than the rest two binary classifiers, which is in accordance with our conjecture in section 3.3. Distinguishing doc/map from photo or painting is easier than distinguish painting from photo. The performances on feature levels $2-5$ only slightly differ in the two easy binary classifiers.

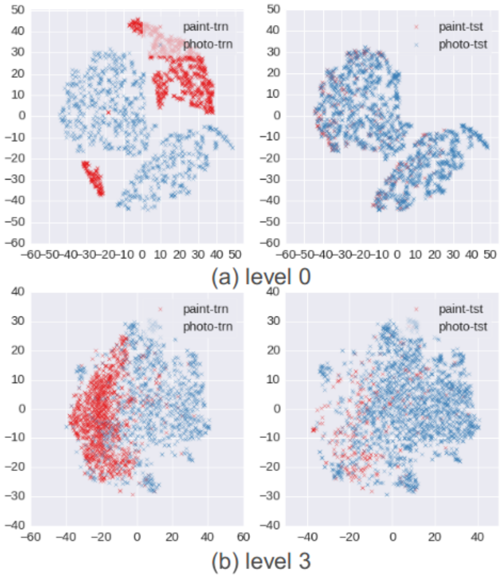
Surprisingly, the level 0 features predict all the images as photos on our history dataset. We visualize level 0 features on the training and test data by using t-SNE algorithm [31] to embed features on a two dimensional plane in Figure 5(a). We can see that the hand-crafted level 0 features designed in work [4] works perfectly in separating paintings from photos on the training data. However, these features are entirely mixed for photos and paintings on the test data. We also visualize the level 3 features in Figure 5(b) as a comparison. From the comparison, we observe two phenomena. First, the embedded points of paintings in the training data are more dense than those in the test data in both figures. Our explanation is that the paintings in the training data are more typical as they are collected from a particular archive website. Second, only level 0 features exhibit dramatic difference on the two datasets. We suspect that the hand-crafted features are over-tuned as all the experiments in [4] were conducted on paintings from the same website.

We also compare the three fusion strategies on the full genre classification in Table 6. We see that both early and late fused predictors perform better than single level-feature predictor, which indicates that different levels of features are complementary. We illustrate some top predicts for each genre in Figure 6.

4.3 Image Grouping

4.3.1 Iconic Image Grouping

We use pairwise P_{pair} and R_{pair} metrics as they have been widely used in iconic image grouping [34][7]. Larger P_{pair} and R_{pairs} mean higher precision and recall. We sample 46 events for iconic image labelling and label 65 ground-truth iconic image groups with 4.43 images in each group on average. We calculate P_{pair} and R_{pair} for each event. For the images that are detected by the algorithm but not labelled in the ground-truth iconic image grouping, we assign each of them to a new different group ID. That is, we don't consider them to be in the same group with any other images.



(a) level 0

(b) level 3

Figure 5: Visualization of level 0 and level 3 features on training (left) and test (right) data of the painting vs. photo classifier

We compare all the four matching algorithms with different levels of features: SIFT + RANSAC on level 0, SIFT + RANSAC + salient region on level 0, SPM on level $1-4$ and cosine on level $5-6$. As shown in Table 7, each method is shown in mean ± std (standard deviation) format based on different events. Each method is tuned to its best performance on matching thresholds. Paired t-test shows that the improvement of SIFT + RANSAC + salient region over SIFT + RANSAC is significant at 0.05 p-value.

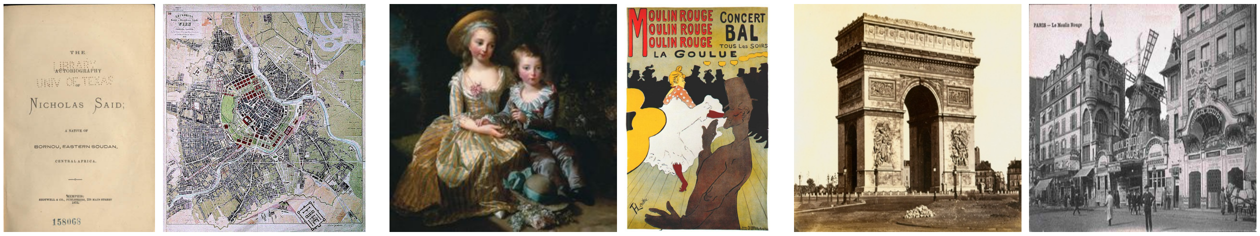
We study four typical modified versions in historic event image profiles on two automatically detected iconic image groups in Figure 7. We select the image with highest resolution as the iconic one on the left and list modified versions in the same iconic image group on the right. The registration of modified versions are highlighted by red boxes in the iconic one.

cropping: a2, a3 and b4 are cropped versions. They don't have higher resolution on the corresponding regions compared to the iconic one. For the cropped version, iconic image grouping can be used to recover the original high resolution one.

editing: a1 seems to be a cropped version but after careful comparison with the iconic one we find that the military on the right is erased from a1. We indicate the position of this military man by a blue box in a1. For the edited version, iconic image grouping can be used to remind users such images may be inconsistent with the historical fact.

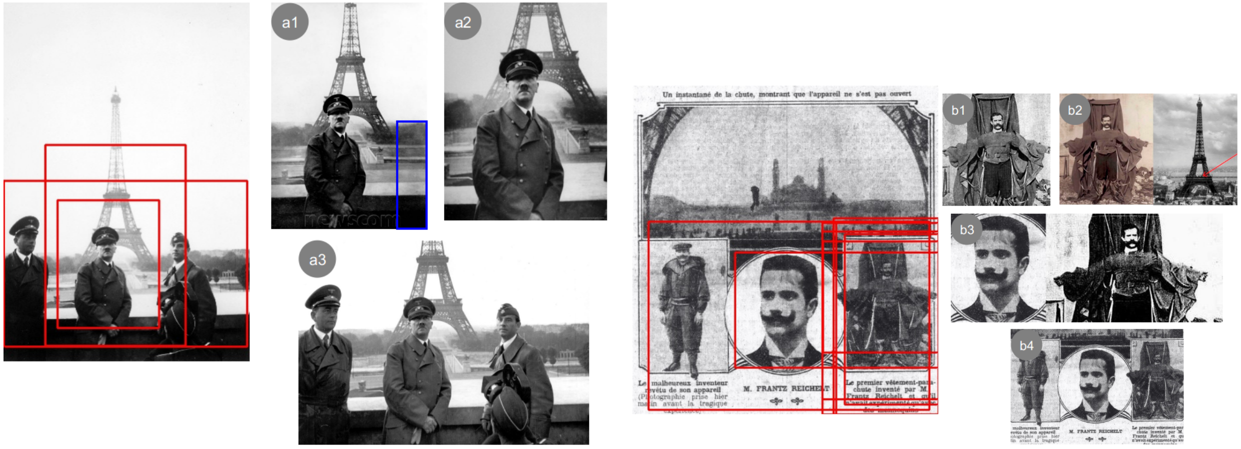
composition: b2 is a manual composition of different history photos. Such composition helps enrich the context of the iconic image. The red arrow in the manually added image illustrates which level of the tower the man jumped from.

higher resolution of detail: b1 and b3 are higher resolution version of corresponding region in the iconic one. They give us a detailed look of the parachute the man wearing.



(a) doc/map (4.8, 2.7) (b) painting (2.3, 1.1) (c) photo (6.1, 4.9)

Figure 6: Top predicts for each genre: the predicted score is shown in the bracket.



(a) (b)

Figure 7: Example of iconic image groups: the selected iconic image is shown on the left and its different modified versions on the web are shown on the right

level	distance	consensus groundtruth				diverse groundtruth			
		ARI	AMI	HM	CM	ARI	AMI	HM	CM
0 (mpeg7)	dist	0.058	0.089	0.327	0.245	0.146	0.188	0.397	0.370
0 (gist)	euclidean	0.056	0.082	0.311	0.246	0.130	0.180	0.364	0.370
$1-4$	SPM	0.118	0.167	0.407	0.323	0.253	0.306	0.481	0.474
$5-6$	cosine	**0.121**	**0.184**	**0.444**	**0.325**	**0.294**	**0.344**	**0.536**	**0.496**

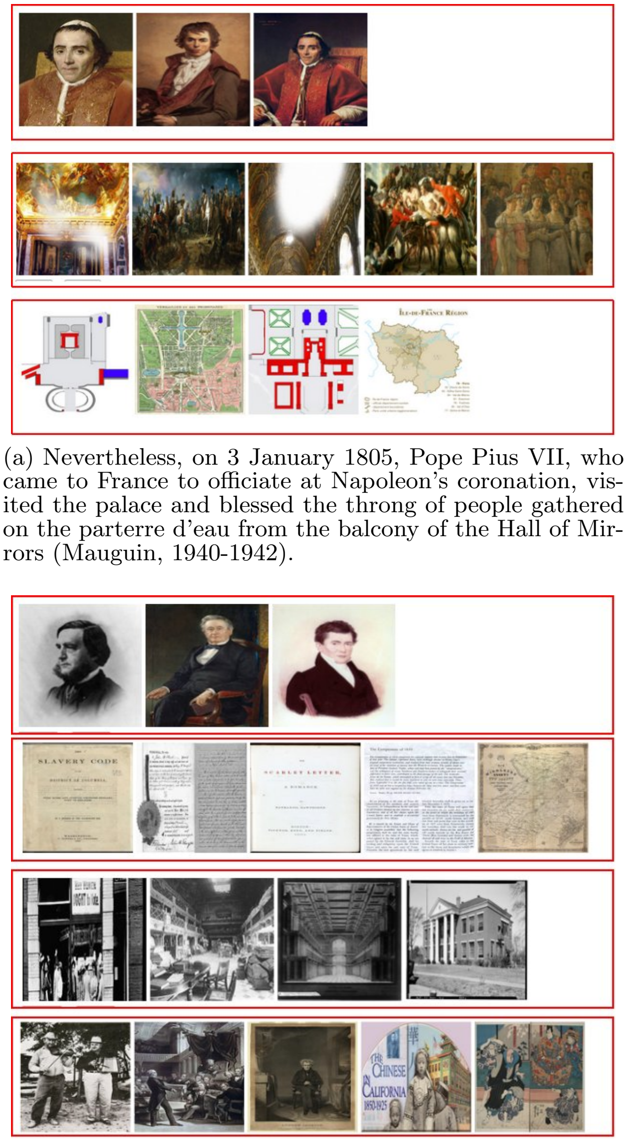
Table 8: Comparison of matching strategies for different feature levels on profile diversification: the parameter in affinity propagation is tuned to its best performance.

4.3.2 Facet Image Grouping

We use ARI [12], AMI [21], HM [26] and CM [26] metrics in evaluating diversified facet image grouping. If the label assignment is entirely random, ARI and AMI values are 0 , which can be used as a baseline value in comparison. ARI and AMI give a normalized (regard to random baseline) single value for evaluation while HM and CM can be used to qualitatively analyze for what "kind" of mistakes are made in the assignment. Facet image grouping is a relatively subjective task and we use mturk to collect the ground-truth, assigning 3 workers per hit. Our instructions for the hit are: 1. At most 5 groups are required. 2. a group should consist of at least 2 images. 3. Not all images are required to be labelled (there may be outliers). For 75 events, we collect 3.84 image groups under each event on average and 6.06 images under each group on average. This indicates the diversity of historic event images.

We note that multiple manual labels for clustering related hit hasn't been well studied yet. Here we propose two ground-truths from multiple labelling of clusters. The first ground-truth is called consensus ground-truth, which includes only images and labels agreed by all the labellers. The second ground-truth is called diverse ground-truth, which comprises three label tracks from three workers respectively.
Details of consensus ground-truth: First we remove images that are labelled by less than 3 workers. On the remaining images, we intersect the groups from all the 3 worker labels. This may result in singular groups which contains only one image and we remove such singular groups. The images in the rest groups are used in evaluation.
Details of diverse ground-truth: The predicted grouping is compared to each track and we report the best evaluated score among the three comparisons. For each track in this ground-truth, we preserve images labelled by the corresponding worker to get evaluation score.

We compare all the proposed different matching strategies on different levels in the full-range features in Table 8. The details of performance change within level $1-4$ and level $5-6$ are not shown due to the limit of space. Within level $1-4$, the best performance is achieved at level 3 while within level $5-6$, the best performance is achieved at level 6. We can see that cosine distance of level 6 performs best on both consensus and diverse ground-truths. Its performance doubles that of level 0 features. This indicates that image profile diversification relies more on content semantics than visual appearance. To further illustrate the semantic coherence within each group, we present case study of image profile diversification on two events in Figure 8. For the event shown at the top, the first group contains portraits



(a) Nevertheless, on 3 January 1805, Pope Pius VII, who came to France to officiate at Napoleon's coronation, visited the palace and blessed the throng of people gathered on the parterre d'eau from the balcony of the Hall of Mirrors (Mauguin, 1940-1942).

(b) During the 1850s the Smithsonian Institution's librarian Charles Coffin Jewett aggressively tried to move that organization towards becoming the United States' national library.

Figure 8: Case study of facet image grouping

for figures in the event; the second group shows the Hall of Mirrors; the third group contains various maps, from indoor balcony and hall map to outdoor Palace of Versailles map. For the event shown at the bottom, the first group shows portraits of the librarian; the second group shows the documents in the library; the third group shows the librarian photos; the fourth group shows the activities and posters of events hosted by the library.

5. CONCLUSION

In this paper, we proposed to automatically construct image profiles for historic events given only a one sentence description with 4Ws. This extremely simple input requirement makes it easy to apply the method to many different historic events and support a wide range of cultural preservation and curation applications ranging from wikipedia enrichment to history education. However, it is also an extremely challenging task which must process relevant information from a wild and unconstrained sources directly. We tackle two major challenges in this work: unconstrained image genres and diversity of image content. Different genres represent different methods of recording history through the image medium. The image content of a historic event can include any type of visual information including faces, objects and scenes, corresponding to a wide range of computer vision tasks. Within an event, images are different in content which reflects different facets of the event. We analyze the data using a full-spectrum of image features including classical hand-crafted low-level features and the rich information in different layers of a convolutional neural network. Comprehensive experiments show the effectiveness of different levels of features from a full-range of features ranging from pixel level to semantic level analysis to solve these challenges. In our future work, we will perform a formal user study and expand the dataset to cover more historic events.

6. ACKNOWLEDGMENTS

This research was partially supported by the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 14XNLQ01), the Beijing Natural Science Foundation (No. 4142029), and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry. This research was also supported in part by the National Science Foundation (NSF) under grant number IIS-1251187. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

7. REFERENCES

[1] B-25 empire state building crash. `http://en.wikipedia.org/wiki/B-25_Empire_State_Building_crash`.

[2] Victor hugo quotes. `http://www.brainyquote.com/quotes/quotes/v/victorhugo385868.html`.

[3] O. Chum and J. Matas. Optimal randomized RANSAC. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(8):1472–1482, 2008.

[4] F. Cutzu, R. I. Hammoud, and A. Leykin. Estimating the photorealism of images: Distinguishing paintings from photographs. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), 16-22 June 2003, Madison, WI, USA*, pages 305–312, 2003.

[5] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255, 2009.

[6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 647–655, 2014.

[7] Z. Feng, J. Chen, X. Wu, and Y. Yu. Aggregation-based probing for large-scale duplicate image detection. In *Web Technologies and Applications - 15th Asia-Pacific Web Conference, APWeb 2013, Sydney, Australia, April 4-6, 2013. Proceedings*, pages 417–428, 2013.

[8] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[9] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587, 2014.

[10] J. Hays and A. A. Efros. Scene completion using millions of photographs. *ACM Trans. Graph.*, 26(3):4, 2007.

[11] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[12] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[13] H. il Koo and N. I. Cho. Rectification of figures and photos in document images using bounding box interface. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3121–3128, 2010.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 1106–1114, 2012.

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 2169–2178, 2006.

[16] X. Li, C. Wu, C. Zach, S. Lazebnik, and J. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I*, pages 427–440, 2008.

[17] M. Lin, Z. Hu, S. Liu, M. Wang, R. Hong, and S. Yan. eheritage of shadow puppetry: creation and manipulation. In *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*, pages 183–192, 2013.

[18] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[19] M. Lux and S. A. Chatzichristofis. Lire: lucene image retrieval: an extensible java CBIR library. In *Proceedings of the 16th International Conference on Multimedia 2008, Vancouver, British Columbia, Canada, October 26-31, 2008*, pages 1085–1088, 2008.

[20] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[21] X. V. Nguyen, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854, 2010.

[22] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[23] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*, 2007.

[24] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2014, Columbus, OH, USA, June 23-28, 2014*, pages 512–519, 2014.

[25] E. Roman-Rangel, C. P. Gayol, J. Odobez, and D. Gatica-Perez. Searching the past: an improved shape descriptor to retrieve maya hieroglyphs. In *Proceedings of the 19th International Conference on Multimedia 2011, Scottsdale, AZ, USA, November 28 - December 1, 2011*, pages 163–172, 2011.

[26] A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 410–420, 2007.

[27] P. Salembier and T. Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, Inc., New York, NY, USA, 2002.

[28] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*.

[29] C. Simon, Williem, J. Choe, I. D. Yun, and I. K. Park. Correcting photometric distortion of document images on a smartphone. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2014, Columbus, OH, USA, June 23-28, 2014*, pages 199–200, 2014.

[30] K. Song, Y. Tian, W. Gao, and T. Huang. Diversifying the image retrieval results. In *Proceedings of the 14th ACM International Conference on Multimedia, Santa Barbara, CA, USA, October 23-27, 2006*, pages 707–710, 2006.

[31] L. van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

[32] R. H. van Leuken, L. G. Pueyo, X. Olivares, and R. van Zwol. Visual diversification of image search results. In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 341–350, 2009.

[33] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. *CoRR*, abs/1412.4564, 2014.

[34] B. Wang, Z. Li, M. Li, and W. Ma. Large-scale duplicate detection for web image search. In *Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, ICME 2006, July 9-12 2006, Toronto, Ontario, Canada*, pages 353–356, 2006.

[35] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3485–3492, 2010.

[36] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 818–833, 2014.

[37] J. Zhang and S. Sclaroff. Saliency detection: A boolean map approach. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 153–160, 2013.

[38] S. Zhao, Y. Gao, X. Jiang, H. Yao, T. Chua, and X. Sun. Exploring principles-of-art features for image emotion recognition. In *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, pages 47–56, 2014.