
SMARTPHONE PICTURE ORGANIZATION: A HIERARCHICAL APPROACH

A PREPRINT

Stefan Lonn

Department of Mathematics and Computer Science
University of Barcelona
Barcelona, Spain

Petia Radeva

Department of Mathematics and Computer Science
University of Barcelona and Computer Vision Center
Barcelona, Spain

Mariella Dimiccoli

Institut de Robòtica i Informàtica Industrial, CSIC-UPC,
Barcelona, Spain.
mdimiccoli@iri.upc.edu

September 13, 2019

ABSTRACT

We live in a society where the large majority of the population has a camera-equipped smartphone. In addition, hard drives and cloud storage are getting cheaper and cheaper, leading to a tremendous growth in stored personal photos. Unlike photo collections captured by a digital camera, which typically are pre-processed by the user who organizes them into event-related folders, smartphone pictures are automatically stored in the cloud. As a consequence, photo collections captured by a smartphone are highly unstructured and because smartphones are ubiquitous, they present a larger variability compared to pictures captured by a digital camera. To solve the need of organizing large smartphone photo collections automatically, we propose here a new methodology for hierarchical photo organization into topics and topic-related categories. Our approach successfully estimates latent topics in the pictures by applying probabilistic Latent Semantic Analysis, and automatically assigns a name to each topic by relying on a lexical database. Topic-related categories are then estimated by using a set of topic-specific Convolutional Neuronal Networks. To validate our approach, we ensemble and make public a large dataset of more than 8,000 smartphone pictures from 40 persons. Experimental results demonstrate major user satisfaction with respect to state of the art solutions in terms of organization.

Keywords smartphone pictures · hierarchical classification · probabilistic latent semantic analysis · convolutional neural networks

1 Introduction

With the proliferation of digital cameras and mobile devices, the number of photos taken each year is growing exponentially. Bolstered by the decrease in price of both hard-drive and cloud storage, people are overwhelmed with their lifetime of photos. The explosive growth of personal photos leads to the problems of photo organization, management and browsing. Indeed, arranging systematically huge photo collections and retrieving specific pictures from them can be a daunting task, which becomes more and more difficult as time passes by ([1, 2]). This has initiated extensive research on content-based image retrieval systems ([3, 4, 5, 6, 7, 8, 9]). Digital photographs typically include metadata in a standard image header, such as time, date and Global Positioning System (GPS) information that can be used for automatic organization. In addition, consumers often organize their photos in directories corresponding to particular “events”, naturally associated with specific times and places such a wedding ceremony or a birthday party.

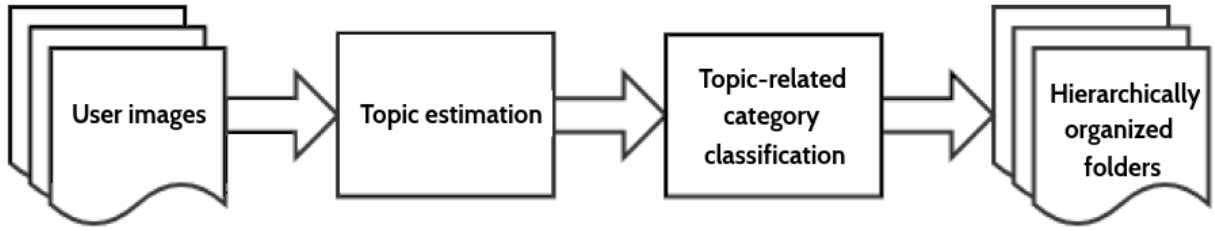


Figure 1: Overview of the proposed approach.

Surprisingly, the organization of pictures captured with a smartphone has received very little attention in the computer vision literature. Smartphone photo collections are in general acquired over a long period of time and typically there is not enough temporal neither semantic structure to be exploited since the pictures can be taken anytime at arbitrarily large interval of time. Beside the lack of structure, the organization of smartphone pictures present additional challenges. Since photos are taken anywhere and anytime and people typically do not regularly remove unwanted/no more useful pictures from the smartphone, cloud-stored pictures include several examples that are in general not observed in a photo collection. Typically, they present a huge variability ranging from notes taken in class to exotic objects seen during a travel on the other side of the ocean. Finally, although the constantly improving smartphone cameras, the quality of pictures due to motion blurring or limited illumination used to be relatively low.

Classifying topics in smartphone photo collections represents an efficient way to organize them. This helps users keep order in their photo collections and also eases the retrieval of similar image types in large photo repositories. Although the problem of smartphone photo organization has attracted the interest of several companies in the market, to the best of our knowledge, there is no work in the computer vision literature that addresses the problem of organizing smartphone pictures. Related work include clustering, segmentation and event classification in photo albums ([10, 11, 12, 13, 14, 3, 4, 5, 15, 6, 7, 9, 8, 16]), photo labelling ([17, 18]), photobook creation ([19, 20, 21]), and event recognition from single images shared online ([22]). However, the approaches proposed so far are not directly applicable to smartphone pictures, since they lack of temporal structure and social network metadata, and present a huge variability in terms of depicted objects, people, scenes, animals and events.

Most of current commercial solutions consist of interactive methods for photo organization, where the definition of the categories and the assignment of a picture to a given category is done manually. Software for automatic photo organization include the popular Eden Photos and Google Photos. Eden Photos provides a coarse classification into a relatively small number of topics, whereas Google Photos provide a finer classification into a large number of categories ranging from abstract concept to concrete objects.

In this work, we propose a more structured classification into a small number of generic topics and a large number of topic-related categories (see Fig.1).

The important benefits of the proposed approach are:

- (i) a hierarchical organization in categories and subcategories instead of state of the art one-level classification solutions,
- (ii) a fully unsupervised approach for category (topic) classification that first discovers latent topic in images and then automatically names them by relying on a lexical database,
- (iii) a very large number of sub-categories for each topic estimated by a set of topic-specific Convolutional Neural Networks (CNN), that are of interest for people who have hobbies, or like to have pictures of a particular topic, and
- (iv) a framework that solely rely on visual data and could be easily enriched with additional information provided by GPS coordinates and EXIF metadata.

As additional contribution, we make public a large subset of our test-set in order to encourage further investigation in the direction of personal smartphone photo organization ¹. User studies demonstrated that the proposed organization achieves better user satisfaction based on experiments performed over a large real-world photo collections.

The reminder of this paper is organized as follows. Section 2 reviews the state of the art on photo organization, while section 3 details the proposed approach. Sections 4.1 and 4.2 describe our experimental setting and discuss the

¹<https://drive.google.com/open?id=1KM0mqudSi6y6HuRaYsBQ3EJbTX1dRrzk>

experimental results, respectively. Finally, section 5 concludes the paper by highlighting the main contributions and outlining future work.

2 Related work

Clustering, segmentation and summarization of photo albums Early algorithms for personal photo organization have mostly relied on temporal and spatial information either to cluster visually similar images into groups while neglecting temporal information or to segment temporally ordered sequences into segments ([10, 11, 12, 13, 14, 15, 16]). More specifically, time metadata, low-level information and, more recently, other picture metadata such as GPS have been used as features for these tasks.

More recently, lifelogs, as particular case of photo albums captured by a wearable photocalera, have attracted lot of research attention ([23]). One of the main challenge is to summarize the huge amount of personal photos collected, with the minimum semantic loss, often according to specific requirements as in the ImageCLEF lifelog summarization task ([24]). In this context, clustering pictures has been proposed as a fundamental step towards summarization. [25] summarized lifelogs by extracting a keyframe from each cluster of images obtained by applying k-means. [26] applied hierarchical clustering to a shortlist of images representing the search query given by the challenge organizers. Then, by relying on the similarity score between each image concept from the cluster and the manually provided topic description, the best image candidates for the lifelog summarization are selected ([27]).

However, clustering, temporal segmentation and summarization are just preliminary step towards photo organization as they can be exploited mostly to support annotation and browsing over a large collection of photos or to assist the creation of photo albums.

Event recognition in photo albums The problem of smartphone picture organization is related to the literature on automatic organization of photo collections and, more in general on image and video event classification. Contrary to video events, photo collections present a very sparse sampling of visual data. Additionally, photo collections are highly ambiguous at a semantic level since many high level features as people for instance are shared across several events. As a consequence, most of the approaches proposed so far, have focused on exploiting the collection structure that is often found in personal and professional photo archives for automatic event classification/image indexing. Typically, such approaches leverage high-level features such as objects, faces, scene, tags ([5, 6, 7, 8]), or time metadata and GPS data ([3, 4, 9]) to automatically label events. For example, [28] exploited prior knowledge about what objects are relevant for a given event in holidays photo collections, to detect events based on object detector outputs. Prior knowledge was obtained statistically from mass image collection web site. [29] proposed a probabilistic fusion framework that integrates the prediction from individual photos to obtain the collection level prediction. The idea of using a fusion framework was later adapted by [30], who proposed a coarse-to-fine hierarchical model to recognize events in personal photo collections. Similarly to [29], they used multiple features including time, objects, and scenes and relied on CNN features based on the Places database ([31]) to train the coarse classifier for coarse event recognition. CNN features for objects and time features are used to train fine classifiers with the three features. Finally, late fusion is used to get the final predictions.

An original approach was taken by [6] who casted photo collections as sequential data and treated sub-events as latent variables associated to each image in an Hidden Markov Models and learned them while training the event classifier. More recently, [8] proposed a probabilistic graphical model to predict the event categories of groups of photos, that relies on high-level visual features such as objects and scenes extracted directly from images by employing a deep learning based approach.

All these works focus on the recognition of a limited set of social events and are not directly applicable to single snapshots captured by smartphone pictures without temporal structure. Furthermore, a good amount of photos captured by a smartphone is not related to social events but captures a huge variability of objects, animals and places.

Photobook creation and management Another problem closely related to smartphone picture organization is the creation of a photobook from a large personal image collection. Although largely investigated since the advent of social networks nearly a decade ago, photobook creation is still an active area of research. Early approaches were characterized by a large degree of user interaction mainly for labelling ([20, 11]), whereas late approaches aimed at minimizing user supervision by providing multiple picture selections. A representative work is the one of [19], who proposed to combine a chronological representation with a thematic representation. The former is obtained by applying a temporal event clustering algorithm ([32]). The latter is derived from the commonality of metadata features, including EXIF metadata and a combination of low level (i.e. color) and high level image features (i.e. faces). Related to photobook management are the problems of photo browsing and photo galleries compression. [21] enabled a multiscale overview of the photo

albums for efficient browsing and searching. The photos were first grouped into clusters and then displayed sequentially on a user controllable time scale. [33] proposed an alternative solution based on treemap for visualization and presented a study about the ideal parameters for constructing these representations. With the goal of compressing photo galleries created by multiple users attending common social events, [34] proposed a coding strategy relying on geometrical and temporal properties, as well as on the visual content. The approach builds on a graph-based optimization scheme to find the correct ordering between images and on a 3D estimation from matched keypoints to assess image similarity.

All these approaches have been proposed in the context of online social networks or web images, where, contrarily to smartphone pictures automatically stored in the cloud, richer sources of metadata and contextual information are available. From a technical point of view none of these methods rely on topic models but is build on classical clustering techniques on several (groups of) features.

Event recognition from images shared online Nowadays, a large number of photos captured by a smartphone or a digital camera are shared on-line. Typically, the shared images are snapshots of special occasions such as birthdays, weddings, or more in general of social events; or they capture news events such as a marathon, a festival, or a natural disaster. Motivated by this trend, [22] addressed the task of recognizing complex events from still images downloaded for the web, with few labeled examples. Their learning framework uses Wikipedia to generate event categories and noisy Flickr tags as initial pool of concepts, from which event-centric phrases are generated using a tweet segmentation algorithm. Finally, each event category is projected onto a word embedding, nearest neighbors are extracted and added to the pool of segmented phrases. The CNN features of images related to each concept are used to train concept classifiers. The concept scores predicted on a given test image are used as final features for event recognition.

Unlike these works that focus on snapshots shared on-line, and are typically limited to social or news-related events, our work aims to organize all personal photos captured by a smartphone in a hierarchical fashion.

Online photo labeling More recent works have focused on indexing photos on the web shared on social networks such as Picasa ², Flickr ³, Facebook ⁴ and Instagram ⁵. These sharing photo communities generate vast amounts of metadata as users interact with their images that have been exploited for multi-label annotation. [17] proposed a graphical model that explicitly accounts for the inter-dependencies between images sharing common properties that go beyond image tags and include text descriptions and comment threads associated with each image. Moreover, the user profile information is stored including their location and their network of friends, groups, galleries, and collections in which each image was stored. To automatically classify images on the web, the work of [18] builds on the observation that images with similar social-network metadata tend to depict similar scenes. Therefore, given an unlabeled image, contextual information from a neighborhood of images similar to the given one and sharing social-network metadata with that one, is exploited for automatic multi-labeling.

Inspired by the Google image search tool, [9] took a more direct image retrieval approach, aiming at producing relevant content for any user-specified textual query. Since typically only a few pictures are annotated with text, they used picture information as time-stamps, GPS locations, and image pixels to correlate with information on the Internet. More specifically, time-stamps are used to correlate with holidays listed in Wikipedia, GPS location to places listed in Wikimapia, and image pixels to indexed photos, with the goal of dealing with the lack of annotations.

However, all these methods rely on the use of network metadata that are not available for smartphone pictures that have not been shared online or are directly oriented to image retrieval instead of image organization.

Table 1: Example of mixed coefficient $P(z|d_{test})$ obtained for an unseen image.

Topic ID Words	0	1	2	3	4	5	6	7
Crowd	0.000	0.000	0.000	0.000	0.383	0.000	0.617	0.000
Ball	0.000	0.000	0.000	0.000	0.999	0.000	0.001	0.000
Team	0.000	0.000	0.000	0.051	0.936	0.000	0.064	0.000
Skill	0.000	0.000	0.000	0.000	0.918	0.000	0.082	0.000
Flag	0.007	0.000	0.000	0.000	0.730	0.000	0.263	0.000
Stadium	0.000	0.000	0.000	0.000	0.913	0.000	0.087	0.000

²<https://picasa.google.com/>

³<https://www.flickr.com/>

⁴<https://www.facebook.com/>

⁵<https://www.instagram.com/>

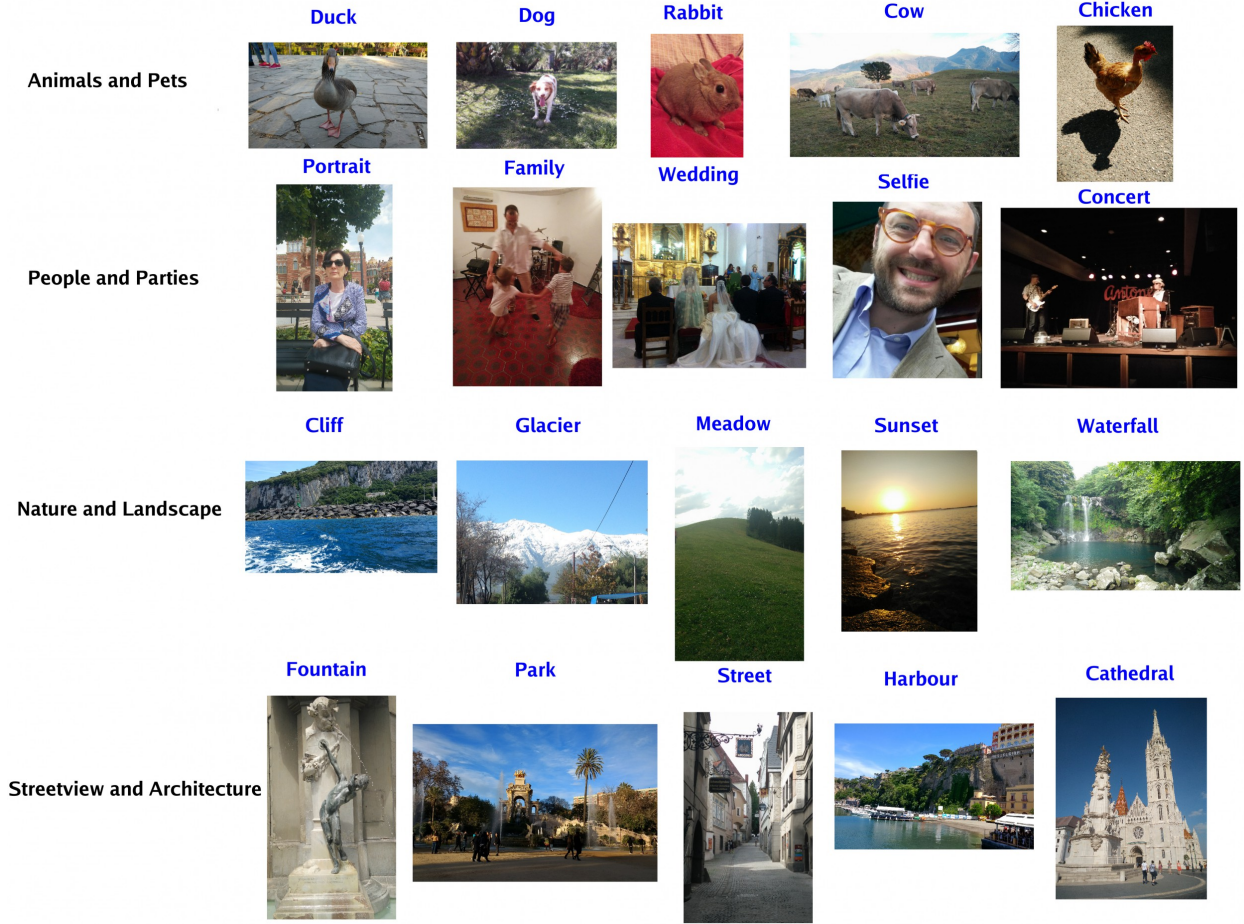


Figure 2: Example of hierarchical organization: each row corresponds to a topic and each column to an example of corresponding category.

Commercial photo organization systems Currently, there are several commercial photo management tools in the market that support photo storage, visualization, labeling, browse, editing, sharing, search and retrieval. Most of them strongly rely on keywords, location, date, person or rating by Exif metadata or annotations. One of the most popular is Google Photos⁶, that automatically arranges uploaded pictures by GPS location and by date taken. Furthermore, it recognizes 1100 different labels, including generic concepts such as *dance* or *kiss*, and objects like *car* or *boots*. However, all this information is grouped into two big categories, *Things*, with 1100 classes and *Places* with a countless number of classes provided by GPS information. While this can be useful for pictures captured during a trip, it becomes less interesting for pictures captured during our daily life since just the name of the city/country is specified. Another widely used software is PicJoy⁷, available on the app Store, that automatically tags your photos by time of day, season, weather, and eventually holiday and provides a visual photo journal. Eden Photos⁸ classifies the user's photos into 14 broad topics, such as *Animals and Pets*, *Text and Visual*. Therefore, photos of *tigers* will appear next to photos of *cats* and *birds*, and photos of *paintings* next to photos of *tickets* or *screenshots*.

Surprisingly, the best organizing software of 2017⁹ such as ACDSee, Zoner photos and PaintShop Pro, does not handle automatic tagging. However, they offer multiple tagging tools and options such as Keywords, descriptions, ratings and labels, GPS tagging using automatic synchronization with tracklogs. Moreover, beside the basic categories *Albums*, *People*, *Places*, and *Various*, new categories are manually added. This kind of interactive solution can be considered good only for photographers who are used to take care of their pictures timely and periodically, not by common

⁶<https://www.blog.google/products/photos/>

⁷<http://www.picjoyapp.com/>

⁸<https://itunes.apple.com/app/eden-photos-heavenly-simple/id1118761521>

⁹<http://www.toptenreviews.com/software/multimedia/best-photo-organizing-software/>

smartphone users who typically have thousands of pictures automatically stored in the cloud and easily forget the pictures they have taken.

Topic modelling in computer vision tasks Although originally conceived for document analysis, topic models have been successfully extended to many computer vision tasks. Initially adapted for object discovery, scene classification, simultaneous classification and segmentation from images ([35, 36, 37, 38]), topic models have been further applied to several video related tasks, including unsupervised learning of human actions ([39, 40]). Typically, each document correspond to an image or video and a codebook representation is learned by performing k-means algorithm on features extracted from each image patch or video shots respectively. Codewords are then defined as the centers of the learned clusters. For the classification task, the latent topic with the highest probability is chosen as the category label of the image. It is worth to stress that during training the image/video categories are known, but the intermediate topic representation, used for testing are learn without additional supervision.

Beside image and video classification and segmentation, topic models have been largely used in the context of image retrieval, but mostly to textual information associated to image data ([41]). For instance, in the context of the MediaEval Retrieving Diverse Social Images Task was required participants to provide the most diverse and relevant images given a search query. [42] proposed to transform image tags and textual description features into a weighted term frequency-inverse document frequency bag-of-words representation ([43]), on the top of which they performed Latent Dirichlet Allocation (LDA) ([44]), to represent topic groups within the results.

A multimodal approach to capture topics in massive social media data has been addressed by [45]. The model, based on the Multimodal-LDA is able to learn correlations between visual and textual modalities.

Unlike classical approaches, where topics are used as an intermediate and more powerful representation for classification that in turn is performed in a supervised manner, in our work topics are the result of a first layer of classification, obtained in fully unsupervised fashion. Furthermore, in all these models the discovery/classification results are given by the topic model itself and none of them uses topic models to drive a more detailed classification as in our system.

In the next section, we detail our proposed approach that provides an automatic hierarchical organization of smartphone pictures such the one shown in Fig.2 by relying solely on visual properties of images. As will be clarified in the next section, the number of topics, their names and the specific topic-related categories have been carefully chosen to address the problem of smartphone picture organization.

3 Proposed approach

Our approach consists of two main steps: topic estimation and topic-related category classification.

3.1 Estimating photo dominant topics

To estimate the dominant topic in an image, we leverage a topic discovery method, called probabilistic Latent Semantic Analysis (pLSA) that has given excellent results in the field of document analysis ([46]). Given a corpus of N documents containing words from a vocabulary of size M , we would like to organize them in K topics.

The corpus of documents is summarized by a $M \times N$ co-occurrence matrix, where each element $X(w_i, d_j)$ with $i = 1, \dots, M, j = 1, \dots, N$ stores the number of occurrence of the word w_i in document d_j . In addition there is a latent variable z_k associated with each occurrence of a word w_i on a document d_j , that represents the topic. The goal of pLSA is to find the topic-specific word distribution $P(w|z)$ and the corresponding document-specific mixing proportions $P(z|d_j)$ which makes up the document specific word distribution $P(w|d_j)$. Formally,

$$P(w|d) = \sum_{k=1}^K P(z_k|d)P(w|z_k) \quad (1)$$

pLSA assumes each document d_j (with word vector w) to be generated from all topics, with document-specific topic weights. The model expresses each document as a convex combination of topic vectors in the latent space with mixture coefficients $P(z_k, d_j)$ for each document d_j , where $k \in \{1, \dots, K\}$. The topic vectors are common to all documents in the corpus and the mixture vectors are specific to each document. For example, in Tab. 1 are shown the mixed coefficients of six words and it can be appreciated how most of these words have the highest coefficient in correspondence of the same topic since it is very likely to find them in the same paragraph of a document.

To learn the topic specific distribution $P(w|z)$ all documents that constitute the training set are pooled together and the PLSA model is fitted to the ensemble of documents for a specified number of topics. In particular, the Expectation

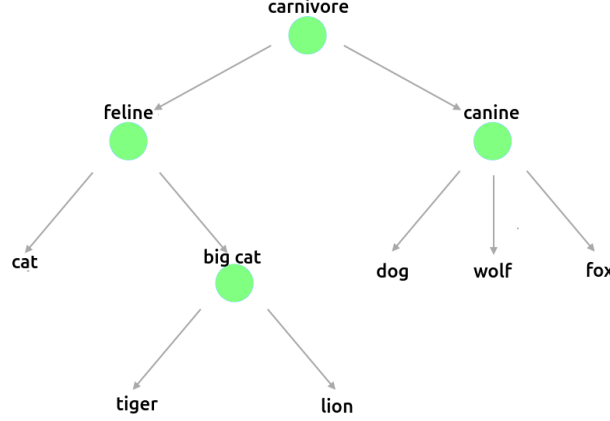


Figure 3: WordNet directed acyclic graph. Each node corresponds to a synset and directed edge from node u to node v indicates that u is an ancestor of v .

Maximization (EM) algorithm ([47]) is used to estimate the parameters $P(z)$, $P(w|z)$ and $P(z|d)$ that maximize the posterior probability $P(z|d, w)$.

Inference and classification Let us suppose that we are given an unseen document, d_{test} and we would like to assign a topic to it. Given the distribution of words in the documents of the test set, say $P(w|d_{test})$, the document specific mixing coefficients, $P(z|d_{test})$ can be computed using the so called *folding-in* heuristic ([48]). When we have a new document d_{test} , the EM algorithm is re-run, but this time the topic-specific word distributions $P(w|z_k)$ are kept fixed to their previous values computed at training, while only the $P(z_k|d_{test})$ are updated. In this way, we obtain the mixed coefficient $P(z|d_{test})$ for the unseen document. The i -th document of test is assigned to the topic k that maximizes the probability of the k -th topic:

$$\operatorname{argmax}_k P(z_k|d_{test}^i), k = 1, \dots, K. \quad (2)$$

Translation in the image domain To adapt this framework to our context, we consider each image as a *document* and each tag, object in the image or concept describing the image as a *word* obtained by applying a concept detector or an object detection algorithm ([49, 50]). In order to apply pLSA, we need first to define a finite vocabulary of words. We build the vocabulary starting by listing all tags that have been used more than 5 times in the training set. This heuristic enforces that all rarely used tags are neglected. If the tag appears only on a few personal photo collections, it is considered rarely used, independently on the count.

Automated topic naming As a result of the inference, we obtain the mixture coefficients that allow to compute the dominant topic of an image with equation (2). We automatically assign a name to the inferred topic k , by using the semantic similarity between the top Q words, (we took $Q = 10$), defining the topic k with highest confidence and $K = 8$ predefined topic names, that we will denote by capitalized words hereafter, namely: *Interior and Objects*, *Pets and Animals*, *Nature and Landscape*, *Food and Drinks*, *Street-view and Architecture*, *People and Portraits*, *Sport and Adventure*, *Text and Visual*. The choice of these topics was inspired by the categories of Eden Photos and motivated by the need of having a small number of categories that could cover all possible content of smartphone pictures. To compute this semantic similarity, we leverage WordNet, a lexical database that groups English words into sets of cognitive synonyms, called *synsets* ([51]). All synsets are connected to other synsets by means of semantic relations. Each vertex v is an integer that represents a synset, and each directed edge (u, v) represents that u is a hypernym (ancestor) of v . The graph is directed and acyclic (see Fig.3). We measure the semantic similarity between two words based on the shortest path in the hypernym taxonomy. Specifically, we used the Lin function ([52, 53]), which is an Information Content (IC)-based similarity measure that relies on the most specific ancestor node, called Lowest Common Subsumer (LCS). Semantically, the LCS represents the commonality of the pair of concepts. For example, the LCS of *mosquito* and *bee* in WordNet is *insect*. If there are multiple candidates for the LCS (due to multiple inheritance), the LCS that results in the shortest path between two input concepts is chosen. Given two synsets, say s_1 and s_2 , their similarity is computed as,

$$S(s_1, s_2) = \frac{2 \cdot IC(LCS(s_1, s_2))}{IC(s_1) + IC(s_2)}. \quad (3)$$

where IC is a measure of specificity for a concept. Higher values are associated with more specific concepts (e.g., chair), while those with lower values are associated to more general concepts (e.g., doctrine). In this work, the IC was derived from *SemCor* ([54]), a manually sense-tagged subset of the Brown Corpus ([55]).

We compute the sum of the Lin similarity between each of the 10 top tags defining the image and the two words in the topic name, for each of the K topics. The topic that has the highest probability is the one that will get assigned to the nameless topic, namely,

$$\operatorname{argmax}_k \sum_{i=1, \dots, Q, j=1, 2} S(s_i, s_j^k), k = 1, \dots, K \quad (4)$$

where s_i is the synset associated to a tag of the image and s_j^k is the synset associated to one of the two words defining the topic k .

3.2 Estimating the topic-related categories

After assigning a topic name to each picture, the proposed method provides a more detailed classification into topic-related category.

For each of our eight topics, we defined the corresponding topic-related categories by relying on topic-related largely used datasets whenever possible. For example, for the topics *Street-view and Architecture* and *Nature and Landscape*, we used the categories of the Places dataset ([31]), that contains 10,624,928 images from 434 categories. We ended up with 277 categories for *Street-view and Architecture* and 88 categories for *Nature and Landscape* respectively. For the *Food and Drinks* topic, we used all 101 categories of the Food101 dataset ([56]). For *Sport and Adventure*, we used the categories of the UCF Sports Action Dataset ([57]) more those relate to Sport of the WIDER dataset ([22]). For *Interior and Object* and *Animal and Pets*, we manually selected the appropriate categories from the ImageNet dataset ([58]) and the Places dataset ([31]). This left us with 428 categories for Interior and Objects and 398 categories for Animals and Pets. For *Text and Visuals*, we defined the categories by inspecting a large training collection of photos captured by a smartphone and identifying images which contained text or some kind of artistic work. Getting specific categories defined is complicated, as many of these categories are defined by the context of the photo instead of the actual content. For example, what differentiates a recipe from class notes is the context and meaning of the text, rather than the visual features which define the image. With this in mind, we defined eleven visual categories which are as follows: *map, screen shot, magazines, drawing, sign, tattoo, poster, graffiti, painting, receipt, writing*. Finally, for *Parties and People*, we defined the following eight categories: *adult, child, selfie, group, family, portrait, manifestation, conference* in addition to 5 categories of the PEC dataset ([6]): *birthday, concert, exhibition, graduation, wedding*. The total number of categories for each topic are detailed on Table 2.

4 Experimental results

In this section, we detail our experimental setting and the experiments performed. Then, we analyze and discuss the results.

4.1 Experimental setting

4.1.1 Dataset

The training dataset was collected with the goal of covering the eight topics defined above. With this goal, we gathered personal photos taken by a mobile phone or a digital camera from 13 subjects having different hobbies (trekking, cooking, traveling, etc), for a total number of 13,845 images, with an average of 1,065 pictures per user. On Table 3, the number of images per user and the number of different topics observed in the pictures are reported.

The test dataset consists of a set of personal photos taken by a mobile phone belonging to 40 subjects, different from those who participated in the collection of the training set, for a total number of 14421 images, with an average of 360 pictures per user.

4.1.2 Validation protocol

We evaluated three different aspects of our proposed approach: 1) how good is the unsupervised classification into topics; 2) how much the users appreciate the proposed hierarchical organization and the appropriateness of the topic and topic-related categories; and 3) the overall classification accuracy of the system.

Table 2: Topic names and number of categories per topics

Eden photos		Hierarchical photo organization	
Topic	#classes	Topic	#classes
Street-view and Architecture	1	Street-view and Architecture	227
Nature and Landscapes	1	Nature and Landscapes	88
People and Portraits	1	People and Portraits	6
Food and Drinks	1	Food and Drinks	101
Text and Visual	1	Text and Visual	11
Animals and Pets	1	Animals and Pets	398
Interior and Objects	1	Interior and Objects	428
Sports and Adventure	1	Sports and Adventure	40
Cars and Vehicles	1	Social events and Parties	12
Macro and Flowers	1	Null	1
Sunrises and Sunsets	1	Google Photos	
Paintings & Art	1	Topic	#classes
Beaches and Seaside	1	Things	1100
Events and Parties	1	Places	Undefined

Table 3: Training dataset composition

Subject ID	1	2	3	4	5	6	7	8	9	10	11	12	13
# Images	527	499	3551	1000	1000	1000	729	1000	1200	823	502	827	1200
# Topics	2	3	5	2	5	3	4	3	2	3	4	3	1

Topic coherence measures To evaluate the performance of a topic model, several topic coherence measures have been proposed that take into account the average or median of pairwise word similarities formed by top words of a given topic. In this work, we used two widely used topic coherence measures: the UCI measure introduced by [59] and the UMass measure introduced by [60]. The UCI-score, \mathcal{C}_{UCI} uses as pairwise score function, the Pointwise Mutual Information (PMI) and is defined as follows:

$$\mathcal{C}_{UCI} = \frac{2}{N(N-1)} \sum_{j=2}^N \sum_{i=1}^{j-1} \log \frac{P(w_i, w_j) + \epsilon}{P(w_i)P(w_j)}, \quad (5)$$

where $P(w_j, w_i)$ is the joint probability of (w_i, w_j) computed as the ratio of number of documents containing both words w_j, w_i , $P(w_i)$ ($P(w_j)$) is the *a priori* probability of w_i (w_j) computed based on frequencies in the dataset, and N is the total number of words. The smoothing count, ϵ is added to avoid calculating the logarithm of zero.

The UMass-score is also based on co-occurrences of word pairs, but measures how much, within the words used to describe a topic, a common word is in average a good predictor for a less common word. More specifically, given an ordered list of words ordered by decreasing frequency $p(w|k)$, say $W = \langle w_1, \dots, w_n \rangle$, it is defined as:

$$\mathcal{C}_{UMass} = \frac{2}{N(N-1)} \sum_{j=2}^N \sum_{i=1}^{j-1} \log \frac{P(w_j, w_i) + \epsilon}{P(w_i)}. \quad (6)$$

Note that the \mathcal{C}_{UMass} has always a negative value.

Additionally, we report the average NMPI (annotated as $AvgNPMI$) among the top Q words as an internal measure of topic coherence:

$$AvgNPMI = \frac{1}{Q(Q-1)} \sum_{j=2}^Q \sum_{i=1}^{j-1} \log \frac{P(w_j, w_k)}{P(w_i, w_k)} \quad (7)$$

where k indicates the k -th topic. The $AvgNPMI$ range is in the interval $\{-1, 1\}$.

Assessing the proposed organization through an user study The proposed approach has been evaluated through an user study, since ultimately the impact of the automatic organization depends on its value to the user. As subjects, we recruit both the 40 owners of the photo collections as well as 30 subjects not involved with the data collection in any way. The photo owners are a valuable resource to discern the photo organization quality, since they only have fully experienced the original content.

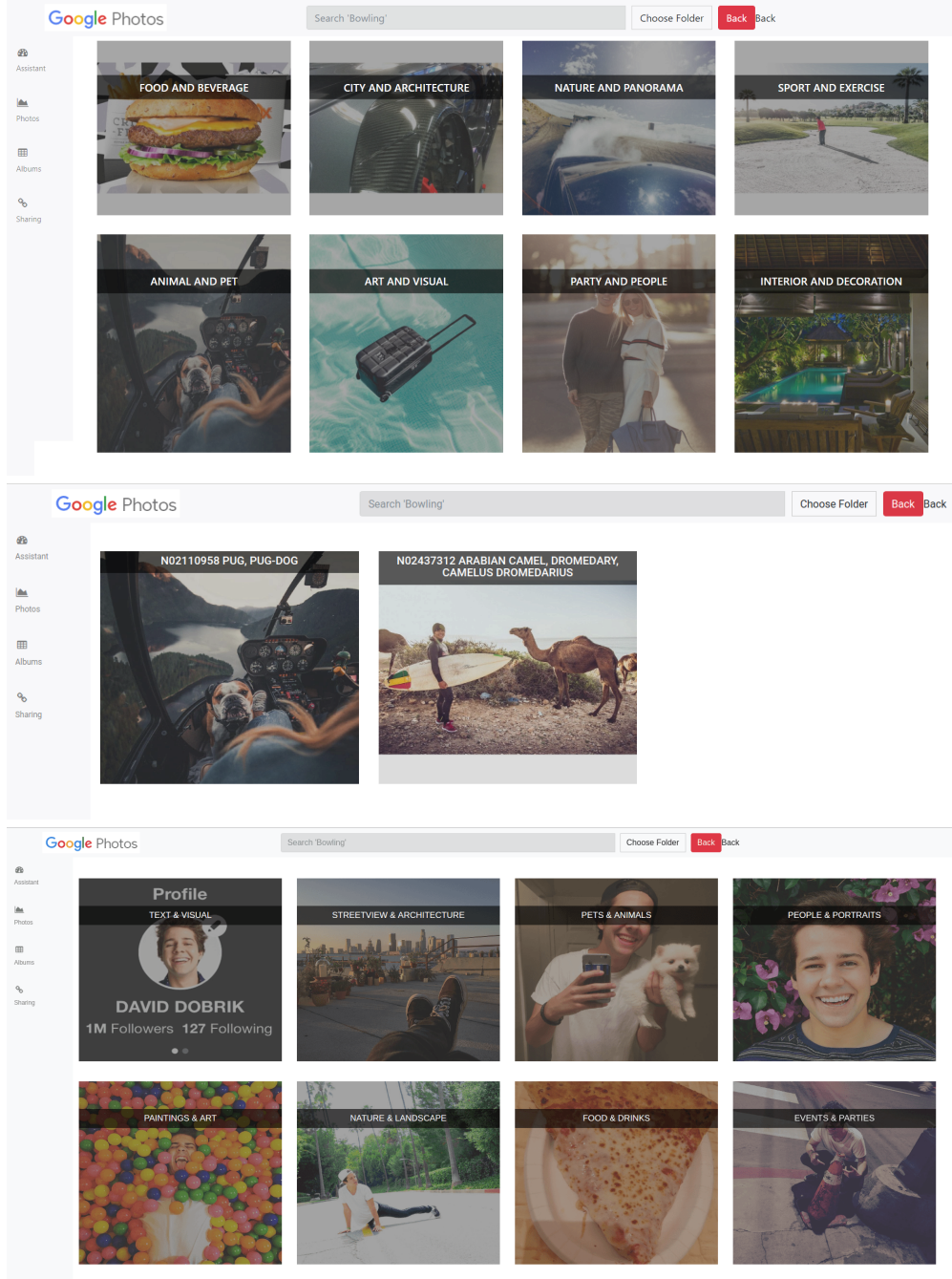


Figure 4: Visual interface used to show to the participants of the user study the results of two different systems. The top two images correspond to the results obtained with our system. In particular, the first image show the categories, and the second one the subcategories of *Nature and Panorama*. The bottom image to the results obtained with Eden photos.

We provided to all participants an Information Sheet that gave them the necessary understanding for the motivation and procedures of the study. To measure the quality of our organization on an absolute scale and to allow independent judges to evaluate the photo organization usefulness, we asked each owner to provide “ground-truth” categories of his/her pictures. Specifically, we asked the users to provide a list of categories that emphasizes the dominant topics in his/her pictures. Then, the users are asked to compare each pair of systems, each one of the pair shown in a different browser tab. To make the systems blind to the users and to avoid bias judgment due to different visualization, we mimicked the visual interface of Google Photos and presented all results using the same interface (see Fig. 4). As shown in Fig. 4,

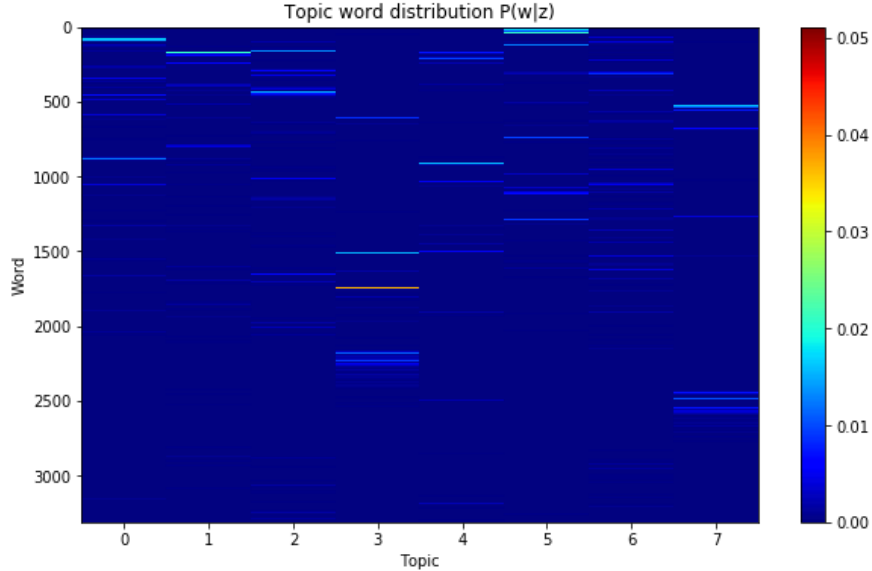
Table 4: Architecture and dataset used for pre-training for each topic

Topic	Architecture	Dataset used for pre-training
Interior and Decoration	ResNet-50	ImageNet
Party and People	ResNet-50	Places
Art and Visual	ResNet-101	ImageNet
Animal and Pet	ResNet-50	ImageNet
Sport and Exercise	ResNet-101	ImageNet
Nature and Panorama	VGG-16	Places365
City and Architecture	VGG-16	Places365
Food and Beverage	ResNet-50	Food-101

the users first see the picture type, and then, by clicking on a picture type they can visualize the subcategories or the pictures corresponding to the picture type. In each subcategory folder, are shown the pictures inside that folder. We asked to the participants two questions: The first question that evaluates aspect a), was: *Which kind of organization do you prefer and why, independently on the accuracy?* The second question, that evaluates aspect b), was: *Which system do provide more accurate results, independently on the organization?*

4.1.3 Experiments

For each user in our test set, we first estimated to which topic the image belongs to and then we classified the image accordingly to the categories of the topic at hand. For instance, if the algorithm predicts that the image belongs to the topic *Animals and Pets*, than a more detailed classification of the pictures is performed with the classes *cats*, *dogs*, *births*, *horses*, *etc.*. In this work, we used a concept detector developed by Imagga Technologies Ltd. Imagga’s auto-tagging technology¹⁰. The advantage of Imagga’s Auto Tagging API is that it can directly recognize over 2,700 different objects and in addition it returns more than 20,000 abstract concepts (corresponding to the words) related to the analyzed images. The total number of tags found in the training set is 13,852. The number of tags after the filtering is 3,312. We then applied pLSA to learn the topic specific word distribution $P(w|z_k)$. At test time, we applied the folding-in heuristic detailed in section 3 keeping $P(w|z_k)$ fixed and we obtained the mixture coefficient $P(z_k|d_{test})$. We automatically assigned a label to the topic with the largest probability. However, if the highest probability is below a given threshold (0.035 in our experiments), the picture is assigned to the *Null* topic.

Figure 5: Topic-specific word distributions $P(w|z_k)$ estimated from the training set.

As it can be appreciated in Table 2, our eight topic categories are a subset of the topic categories in Eden Photos. This is because our system allows several categories for each topic, so that the *Sunrises and Sunset*, *Beaches and Seaside*

¹⁰<http://www.imagga.com/solutions/auto-tagging.html>

Table 5: Comparison of topic models (pLSA, LDA, LSA) in terms of topic coherence measures

Topic	pLSA			LDA			LSA		
	UCI-score	Umass-score	NPMI Topic	UCI-score	Umass-score	NPMI Topic	UCI-score	Umass-score	NPMI
Interior and Decoration	1.40	-1.65	0.16	-	-	-	-0.94 (-0.79)	-2.80 (-2.60)	-0.01 (0.00)
Party and People	1.54	-1.72	0.16	0.87	-0.80	0.20	-	-	-
Art and Visual	1.60	-1.58	0.20	-1.10 (-4.63)	-3.76 (-9.36)	-0.05 (0.03)	3.39	-0.59	0.52
Animal and Pet	1.65	-1.87	0.16	-7.27	-11.43	-0.13	2.31 (1.98)	-1.10 (-1.34)	0.39 (0.30)
Sport and Exercise	1.41	-1.80	0.13	-	-	-	-1.83	-5.04	0.13
Nature and Panorama	1.76	-1.69	0.21	-5.28	-9.38	-0.05	-	-	-
City and Architecture	1.41	-1.69	0.15	1.06	-0.85	0.22	1.12	-0.58 (-5.04)	0.23
Food and Beverage	1.44	-1.78	0.14	-1.12 (-4.10)	-0.19 (-3.88)	-0.05 (0.04)	-0.37	-0.12	0.06
<i>Average Topic Coherence</i>	1.53	-1.72	0.16	-2.69	-4.95	0.02	0.6	-2.23	0.20

and *Flowers* can be considered as categories of *Nature and Landscape* instead of being a topic itself. Similarly, we treated *Painting and Art* a category of *Text and Visual* and *Cars and Vehicles* as a subcategory of *City and Architecture*.

Table 6: Results of the user study based comparison of our system vs Google Photos (top) and our system vs Eden Photos (bottom) on the dataset consisting of 40 users. Numbers indicate percentage of responses for each question.

		Much better (5)	Better (4)	Similar (3)	Worse (2)	Much worse (1)	Mean	Std	Up pvalue	ICC1
Photos owners	Organization	55%	42.5%	2.5%	0%	0%	4.55	0.55	1	-
	Accuracy	0%	20%	40%	37.5%	2.5%	2.77	0.80	0.04	-
External evaluators	Organization	48.34%	40.83%	10.00%	0.83%	0%	4.36	0.69	1	0.430
	Accuracy	1.66%	16.67%	47.5%	31.67%	2.5%	2.84	0.78	0.01	0.436

		Much better (5)	Better (4)	Similar (3)	Worse (2)	Much worse (1)	Mean	Std	Up pvalue	ICC1
Photos owners	Organization	40%	55%	5%	0%	0%	4.35	0.57	1	-
	Accuracy	0%	25%	55%	20%	0%	3.05	0.67	0.67	-
External evaluators	Organization	19.66%	63.34%	16.67%	0.84%	0%	4.07	0.65	1	0.402
	Accuracy	4.16%	19.16%	50.84%	24.17%	1.67%	2.93	0.86	0.20	0.403

4.2 Results and discussion

In the following, we report and discuss the results obtained for topic discovery and assignment, as well as the results of the user study.

Once classified into topics, the images were fed to the corresponding CNN that classified them into topic-related categories. A description of the CNN architectures used for each topic and the initial weights used are provided in Table 4. In order to build the training dataset for fine-tuning, we needed a large amount of photos, ideally taken with a smartphone, as these are impromptu ones, that can be blurred or lacking proper lightening or having the motif of the photo off-centered. With the goal of getting a large amount of smartphone pictures, we scraped social media, such as Instagram and Flickr, and we also got additional photos from Google Images when needed. We automatically collected a large amount of photos per category, and later we manually filtered the ones that did not fit our criteria. Our goal was to get at least a thousand images per category to be able to fine-tune a pre-trained CNN.

4.2.1 Topic discovery

After fitting the pLSA model to our training set with 8 topics and automatically assigning a label to each word distribution, we obtained the following topic definitions:

- *Food and Beverages*: fresh, healthy, dinner, eating, plate, meal, restaurant, delicious, diet, lunch, tasty, gourmet, snack, cuisine, nutrition, dish, vegetable, meat, cook, breakfast, pepper, sauce, tomato, vegetables, slice, kitchen, hot, cheese, bread, bowl.
- *Animals and Pets*: animal, dog, canine, domestic animal, pet, mammal, domestic, person, hunting dog, fur, cat, animals, funny, pets, adorable, sporting dog, purebred, terrier, feline, puppy, breed, hound, furry, kitten, eye, toy dog, spaniel, fluffy, little, whiskers.
- *Art and Visual*: paper, element, shape, text, frame, drawing, money, card, flower, letter, blank, internet, representation, decorative, curve, currency, note, artistic, sketch, surface, document, book, floral, swirl, textured, leaf, information, creative, word, writing.

Table 7: Results of the user study based comparison of our system vs Google Photos (top) and our system vs Eden Photos (bottom) on the dataset consisting of 5 Instagram’s vloggers. Numbers indicate percentage of responses for each question. The top rows report the evaluation made by the photo-owner, whereas the bottom rows refer to the average evaluation made by three users that saw the pictures for the first time.

		Much better	Better	Similar	Worse	Much worse
Google Photos	Organization	55%	42%	3%	0%	0%
	Accuracy	0%	20%	40%	37.5%	2.5%
		Much better	Better	Similar	Worse	Much worse
Eden Photos	Organization	41%	55%	14%	0%	0%
	Accuracy	0%	30%	20%	50%	0%



Figure 6: Example of images correctly classified by our system

- *Nature and Panorama*: europe, rocks, shoreline, barrier, surf, seaside, asia, boundary, sunrise, hill, sunshine, seashore, ship, vessel, evening, sandbar, structure, rocky, peace, coastal, geological formation, turquoise, natural elevation, cloudscape, dusk, pacific, cliff, panorama, scenics, breakwater.
- *Parties and People*: person, caucasian, boy, couple, together, girls, clothing, indoors, family, friends, teenager, 20s, two, group, standing, friendship, working, laughing, blond, brunette, teen, student, romance, education, kid, adults, relationship, mother, romantic, healthy.
- *Sport and Exercise*: person, caucasian, boy, active, healthy, family, kid, playing, play, athlete, childhood, ball, children, player, activity, team, game, two, exercise, little, training, soccer, football, baby, mother, fitness, toddler, match, adorable, care.
- *City and Architecture*: group, crowd, spectator, town, pedestrian, buildings, stage, event, transportation, dark, vehicle, meadow, skyline, high, panorama, snow, music, center, aerial, wheeled vehicle, party, entertainment, tower, countryside, disco, club, transport, power, dance, concert.
- *Interior and Decoration*: wall, window, structure, interior, wood, door, furniture, luxury, estate, apartment, living, exterior, decor, sofa, residential, real, indoors, religion, comfortable, monument, tower, town, famous, church, inside, couch, lamp, historical, brick, europe.

Fig. 5 plots the topic-specific word distributions $P(w|z)$ computed on the training set and shows how different words have different probabilities of appearing in each topic. It can be observed that the three topic measures show consistent results for pLSA. The topic *Sport and Exercise* is the less coherent whereas *Nature and Panorama* and *Art and visual*

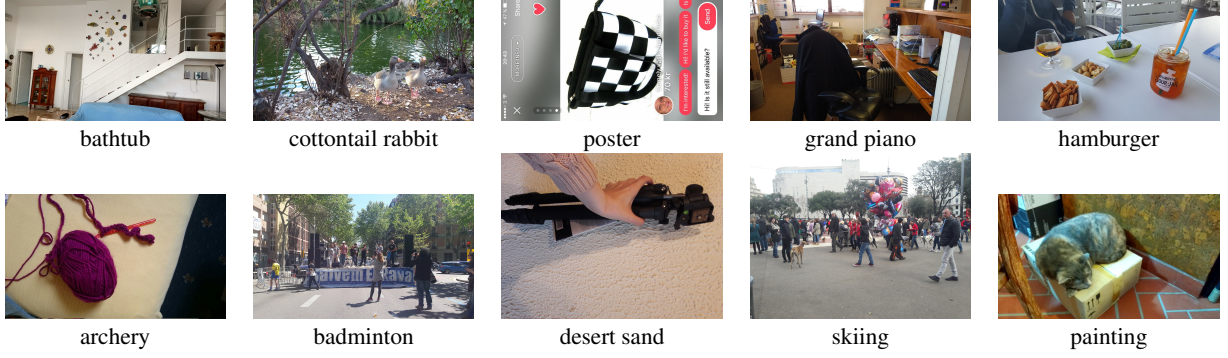


Figure 7: Example of images incorrectly classified: correct topic, but incorrect category (top), incorrect topic and incorrect category (bottom)

Table 8: Datasets and event classes used by state of the art algorithms

Holiday 1 [8, 30]	[7] [5]	Holiday 2 [4, 3]	SocEID [28]	[29]	UIUC Sports [22]
Mardi gras	Beach fun	Christmas	Birthdays	Christmas	Rowing
Thanksgiving	Graduation	Halloween	Graduations	Halloween	Badminton
Christmas	Urban tour	Easter	Marathons/Races	Valentines	Polo
Memorial day	Yardprk	Thanksgiving	Weddings	4 July	Bocce
New Years's Eve	Ball games	Independence's day	Protests	Outdoor Sports	Snowboarding
Easter	Birthday	New Years's Eve	Parades	Birthday	Croquet
Valentine's day	Christmas	Mardi gras	Soccer's matches	Beach	Sailing
Independence's day	Family time	Memorial day	Concerts	Null event	Rock climbing
Halloween	Eating	San Patrick's day			Baseball
San Patrick's day	Skiing	Valentine's day			
	Wedding	Labor day			
	Null event	Mother's day			
PEC [6, 8]	Rare Event Dataset [22]		WIDER [22]		
Birthday	J. Trudeau elected	Engagement parties	Parade	Soldier drilling	Photographers
Children birthday	Election Trump	Boston red sox wins	Handshaking	Spa	Raid
Christmas	Hurricane Katrina	Humanity washed ashore	Demonstration	Sports fan	Rescue
Concert	Hurricane Sandy	Hot air balloon	Riot	Students Schoolkids	sports-Coach-Trainer
Boat cruise	Nepal earthquake	Israel-Palestine conflict	Dancing	Surgeons	Voter
Easter	2012 summer Olympics	Mali attacks	Car accident	Waiter- Waitress	Angler
Exhibition	Obama wins elections	Paris attacks	Funeral	Worker-Laborer	Hockey
Graduation	Columbia space shuttle disaster	Royal wedding	Cheering	Running	People driving car
Halloween	Arab spring	Yemen civil war	Election campaign	Street battle	Traffic
San Patrick's day	9/11 attacks	Thanksgiving	Russian airlines crashes Sinai	Press conference	Basketball
Road trip	Boston bombing	US invasion Afghanistan	People marching	Football	Interview
Hiking	Russian airstrikes Siria	Meeting	Soccer	Group	Celebration or party
Skiing	Baby showers		Tennis	Meeting	Dresses
Wedding	Drones attacks Yemen Pakistan		Ice skating	Greeting	Parachutist paratrooper
			Gymnastic	Ballooning	Aerobics
			Swimming	Car racing	

are the most coherent. For comparison purpose, in Table 5, we report the values of the topic coherence measures described in section 4.1.2, for two other widely used topic models, namely Latent Dirichlet allocation (LDA) [44] and Latent Semantic Analysis (LSA) [61]. These results were obtained by using the Movie corpus, a Wikipedia subset, as external corpus in the *gensim* Python library. As it can be observed, the average topic coherence in terms of UCI-score and Umass-score is much higher for pLSA than LSA and LDA, whereas the NPMI is slightly better for LSA. However, only three topics out of eight have a higher NPMI values, while most of them have very low values. Furthermore, while with the results of pLSA each word distribution was automatically assigned to a different label, with the results of LDA and LSA two different word distributions were assigned to a same label. In particular, with LDA *Art and Visual* and *Food and Beverage* were assigned twice to a same word distribution and the labels *Sport and Exercise* and *Interior and Decoration* were not assigned to any word distributions. Similarly, with LSA, the labels *Interior and Decoration* and *City and Architecture* were assigned twice to a same word distribution, whereas *Party and People* and *Nature and Panorama* were not assigned to any word distribution.

4.2.2 User study results

We recruited thirty persons for the user study who were not involved with the data collection, and six of them were computer illiterate. In average, each photo collection has been evaluated by three different participants. The evaluations were slightly harsher depending on the participant background. We observed that people familiar with technology gave more feedback. Each participant evaluated at most three photos collections. First, we asked people to draw down the

categories into which they would like to organize their pictures. The most popular categories were: *Friends, Architecture, Travel, Panorama, Selfies, Food, Documents, Dogs, Sport* (described with the favorite one such as Skatering). Less common categories were often related to the user job or to a particular hobby.

We compared our photo organization to the two most popular and automatic photo categorization systems, namely Eden and Google Photos. We evaluated two important aspects: a) categories organization, that is hierarchical organization versus just one layer classification, and b) image assignment to the categories. Regarding a), note that Eden has only 14 generic categories, whereas Google has 1100 subcategories. Our system has 8 generic topics and a total of 1311 subcategories (see Table 2).

In Table 6, we report the results of the user study, together with the corresponding statistical descriptors. Specifically, we applied a One Sample T-Test, whose null hypothesis was that the proposed system is "better" or "much better" than the control system in terms of organization or accuracy. To the rates from "much worst" to "much better" we assigned scores from 1 to 5. An up-value larger or equal than 0.04 indicates that null hypothesis is true. Therefore, in terms of organization, our system is considered better than the two control systems, both by the photo owners and external evaluators in a statistical relevant way. We also observed that photo owners gave score slightly higher than external evaluators. In terms of accuracy, the null hypothesis is rejected by both the external evaluators and the photo owners when comparing to the accuracy of Google. However, the value of the mean is very close to the similarity value, that is 3, in both cases (2.77 and 2.84). With Eden Photos, we observed a different trend: the photo owners judged our system having slightly better accuracy than Eden and this result is statistically relevant, whereas external participants considered the accuracy of Eden slightly better (mean value 2.93) but the null hypothesis cannot be rejected since the up pvalue is larger than 0.04. However, when analysing these results it must be taken into account that Google Photo classified in average 53.66% of the pictures, whereas the Eden Photos app 61.29% and our system 81.6%. Indeed, we used an accuracy threshold only for the classification into topics so that all pictures fed to neural networks were considered in the user study without taking into account the classification confidence.

To evaluate the reliability of user study results, we computed the Intraclass Correlation Coefficient (ICC) one-way random commonly known as ICC1 ([62]), since the raters who rate one user were not necessarily the same as those who rate another user. This design corresponds to a One-way Analysis of Variance (ANOVA) in which User is a random effect, and Rater is viewed as measurement error. The ICC1 is a measure of absolute agreement and is sensitive to difference in means between raters. It is defined as follows:

$$ICC1 = \frac{BMS + WMS}{BMS(k - 1)WMS'}, \quad (8)$$

where k is the number of judges rating each target, (BMS) is a between-targets mean square and (WMS) a within-target mean square. As shown in Table 6, we obtained values above 0.4 when comparing both the organization and the accuracy to those of control systems, which is considered a good value for ICC1 ([63]), since ICC1 values are always lower than other intraclass coefficient measures not applicable to this context. These results were obtained by using the ICC package in R ¹¹.

It is very important to remark that Google Photos always classifies the images into a relatively small set of categories, in average 9 over the 40 users, although it is supposed to account for 1100 categories. Furthermore, several participants observed that many categories such as *sky, flowers* or *car* include all pictures where even a small portion of sky (or a car or a flower in the background) is visible and therefore were judged ambiguous. Several other groups of categories such as *food, cooking, recipes* and *baking* were judged redundant. The same occurs for *skyscrapers, skylines, towers*. Furthermore, the only category related to people that was found in the full testing set was *selfies*. A number of participants commented that the categories of our proposed system better reflect the way they would organize their own pictures. However, some participant commented that it would be useful for our system to have intermediate categories. For instance between *Animal and Pets* and *irish terrier*, it would be useful to have the intermediate category *dog*. Although we did not show this in our user study, it is worth to observe that such intermediate classes are naturally provided by the synset associated to the subcategories. Additionally, other participants commented that it would be better to have the opportunity to choice for which topics to have subcategories and for which not. Others commented that sometimes the number of categories is too large given the number of pictures under the topic. These remarks suggest that the number of subcategories could be determined depending on the number of pictures under the main topic, that often unveils what the user like to capture with his/her smartphone. Finally, some participants commented that having information about the place is very important. We stress that many of the user suggestions could be easily integrated into the proposed approach by relying on additional information such as gps coordinates and EXIF metadata. Overall, beside validating the proposed approach, the information collected through the user study will be useful for future developments.

¹¹<https://cran.r-project.org/web/packages/ICC/index.html>

4.2.3 Qualitative results

Fig. 6 and 7 show examples of pictures correctly and incorrectly classified by our system, respectively. In particular, in Fig. 6 it is possible to appreciate the level of detail that can be achieved by our system. On the first row of Fig. 7, are shown examples of pictures that have been assigned to the right topic, but to the incorrect category, whereas on the second row are shown examples of pictures that have been assigned to a wrong topic. Since both topics *People and Portraits* and *Sports and Adventure* involve people, pictures with crowd are easily wrongly assigned to *Sports and Adventure*.

4.2.4 Discussion

Existing algorithms for event recognition from personal photo collections have focused on the detection of a limited set of social events (see Table 8). Even if the PEC dataset becomes a standard in the community, several other in-house datasets with very similar categories (see top part of Table 8) have been used in the literature ([3, 4, 28, 29, 28, 7, 22]). However, in smartphone photo collections there are very few images that have been captured during the same event during a short period of time. Additionally, images captured by a smartphone have a large variability in terms of topics, so that those belonging to the category *Parties and People* are just a (small) portion of them. For these reasons, a comparison with such methods would be unfair.

Although our work is closely related to image categorization for easing image access (seek, organize and understand images), it could serve also as basis for image retrieval. Indeed, since a multilabel approach is typically best suited for retrieval, the same image tags that we used as input to our system could be used together with the hierarchical labels to improve retrieval performance. In addition, time and localization metadata are easily available on a smartphone through GPS and Google Calendar. As demonstrated by the literature on event recognition, the use of this information would be extremely useful in detecting special events. For instance, this would help to retrieve and recognize rare events such the ones of the Rare Event Dataset (see Table 8), that are currently not handled by the proposed system. We leave this for future work.

5 Conclusions

This paper addressed the problem of organizing smartphone pictures into a set of topics and topic-related categories. The proposed approach first classifies images into eight topics by using an unsupervised generative approach that allows to account for their huge intra-class variability. Next, pictures are classified into a large number of categories by using a CNN approach.

User studies demonstrated that users prefer our two-levels classification with respect to a one-level classification provided by widely used photo organization systems such as Eden Photos and Google Photos. The proposed approach could be easily integrated in a retrieval system that relies on both semantic tags, time and location metadata to retrieve all images corresponding to the user query. With the goal of encouraging further research on smartphone picture organization, we make available a dataset of smartphone pictures from 40 persons.

Acknowledgments

We kindly acknowledge the European Network on Integrating Vision and Language (iV&L Net) ICT COST Action IC1307 for the support of a Short Term Visit of one of the authors to Imagga Ltd. The authors are very grateful to Stavri Nikolov for inspiring conversations. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used in our research. Finally, we thank two anonymous reviewers for their helpful and constructive comments that greatly contributed to improving the final version of the paper.

This work was partially funded by TIN2015-66951-C2-1R, SGR-1219, Grant 20141510 (Marató TV3), CERCA Programme / Generalitat de Catalunya and ICREA Academia grant.

References

- [1] Nancy A Van House. Collocated photo sharing, story-telling, and the performance of self. *International Journal of Human-Computer Studies*, 67(12):1073–1086, 2009.
- [2] Steve Whittaker, Ofer Bergman, and Paul Clough. Easy on that trigger dad: a study of long term family photo retrieval. *Personal and Ubiquitous Computing*, 14(1):31–43, 2010.

- [3] Liangliang Cao, Jiebo Luo, and Thomas S Huang. Annotating photo collections by label propagation according to multiple similarity cues. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 121–130. ACM, 2008.
- [4] Liangliang Cao, Jiebo Luo, Henry Kautz, and Thomas S Huang. Image annotation within the context of personal photo collections using hierarchical event and scene models. *IEEE Transactions on Multimedia*, 11(2):208–219, 2009.
- [5] Shen-Fu Tsai, Liangliang Cao, Feng Tang, and Thomas S Huang. Compositional object pattern: a new model for album event recognition. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1361–1364. ACM, 2011.
- [6] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Event recognition in photo collections with a stopwatch hmm. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1193–1200, 2013.
- [7] Jia-Min Gu, Yi-Leh Wu, Wei-Chih Hung, and Cheng-Yuan Tang. Personal photo organization using event annotation. In *Information, Communications and Signal Processing (ICICS) 2013 9th International Conference on*, pages 1–4. IEEE, 2013.
- [8] Siham Bacha, Mohand Saïd Allili, and Nadjia Benblidia. Event recognition in photo albums using probabilistic graphical models and feature relevance. *Journal of Visual Communication and Image Representation*, 40:546–558, 2016.
- [9] Neeraj Kumar and Steve Seitz. Photo recall: Using the internet to label your photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 771–778, 2014.
- [10] Yanfeng Sun, Hongjiang Zhang, Lei Zhang, and Mingjing Li. Myphotos: a system for home photo management and processing. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 81–82. ACM, 2002.
- [11] Khalid Latif, Khabib Mustofa, and A Min Tjoa. An approach for a personal information management system for photos of a lifetime by exploiting semantics. In *International Conference on Database and Expert Systems Applications*, pages 467–477. Springer, 2006.
- [12] Ming Zhao, Yong Wei Teo, Siliang Liu, Tat-Seng Chua, and Ramesh Jain. Automatic person annotation of family photo album. In *International Conference on Image and Video Retrieval*, pages 163–172. Springer, 2006.
- [13] Hyunmo Kang, Benjamin B Bederson, and Bongwon Suh. Capture, annotate, browse, find, share: Novel interfaces for personal photo management. *International Journal of Human- Computer Interaction*, 23(3):315–337, 2007.
- [14] Windson Viana, Jose Bringel Filho, Jerome Gensel, Marlene Villanova-Oliver, and Herve Martin. Photomap: from location and time to context-aware photo annotations. *Journal of Location Based Services*, 2(3):211–235, 2008.
- [15] Fuming Sun, Haojie Li, and Xueming Wang. Photo 4w: Mobile photo management on what, where, who and when. *Neurocomputing*, 119:59–64, 2013.
- [16] Nuno Datia, João Moura Pires, and Nuno Correia. Time and space for segmenting personal photo sets. *Multimedia Tools and Applications*, 76(5):7141–7173, 2017.
- [17] Julian McAuley and Jure Leskovec. Image labeling on a network: using social-network metadata for image classification. *Computer Vision–ECCV 2012*, pages 828–841, 2012.
- [18] Justin Johnson, Lamberto Ballan, and Li Fei-Fei. Love the neighbors: Image annotation by exploiting image metadata. In *Proceedings of the IEEE international conference on computer vision*, pages 4624–4632, 2015.
- [19] Mark D Wood, Madirakshi Das, Peter O Stubler, and Alexander C Loui. Event-enabled intelligent asset selection and grouping for photobook creation. *Image and Vision Computing*, 53:57–67, 2016.
- [20] Yuli Gao, Clayton Brian Atkins, Phil Cheattle, Jun Xiao, Xuemei Zhang, Hui Chao, Peng Wu, Daniel Tretter, David Slatter, Andrew Carter, et al. Magicphotobook: designer inspired, user perfected photo albums. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 979–980. ACM, 2009.
- [21] Kolbeinn Karlsson, Wei Jiang, and Dong-Qing Zhang. Mobile photo album management with multiscale timeline. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1061–1064. ACM, 2014.
- [22] Unaiza Ahsan, Chen Sun, James Hays, and Irfan Essa. Complex event recognition from images with few training examples. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 669–678. IEEE, 2017.
- [23] Marc Bolanos, Mariella Dimiccoli, and Petia Radeva. Toward storytelling from visual lifelogging: An overview. *IEEE Transactions on Human-Machine Systems*, 47(1):77–90, 2017.

- [24] Bogdan Ionescu, Henning Müller, Mauricio Villegas, Helbert Arenas, Giulia Boato, Duc-Tien Dang-Nguyen, Yashin Dicente Cid, Carsten Eickhoff, Alba G Seco de Herrera, Cathal Gurrin, et al. Overview of imageclef 2017: Information extraction from images. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 315–337. Springer, 2017.
- [25] A Garcia del Molino, B Mandal, J Lin, J Hwee Lim, V Subbaraju, and V Chandrasekhar. Vc-i2r@ imageclef2017: Ensemble of deep learned features for lifelog video summarization. In *proceedings of CLEF*, 2017.
- [26] Mihai Dogariu and Bogdan Ionescu. A textual filtering of hog-based hierarchical clustering of lifelog data. *CLEF working notes, CEUR (September 11-14 2017)*, 2017.
- [27] Bogdan Ionescu, Alexandru-Lucian Gînsca, Bogdan Boteanu, Adrian Popescu, Mihai Lupu, and Henning Müller. Retrieving diverse social images at mediaeval 2014: Challenge, dataset and evaluation. In *MediaEval*, 2014.
- [28] Shen-Fu Tsai, Thomas S Huang, and Feng Tang. Album-based object-centric event recognition. In *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, pages 1–6. IEEE, 2011.
- [29] Feng Tang, Daniel R Tretter, and Chris Willis. Event classification for personal photo collections. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 877–880. IEEE, 2011.
- [30] Cong Guo and Xinmei Tian. Event recognition in personal photo collections using hierarchical model and multiple features. In *Multimedia Signal Processing (MMSP), 2015 IEEE 17th International Workshop on*, pages 1–6. IEEE, 2015.
- [31] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- [32] Alexander C Loui and Andreas Savakis. Automated event clustering and quality screening of consumer pictures for digital albuming. *IEEE Transactions on Multimedia*, 5(3):390–402, 2003.
- [33] John A Guerra-Gomez, Cati Boulanger, Sanjay Kairam, and David A Shamma. Identifying best practices for visualizing photo statistics and galleries using treemaps. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 60–63. ACM, 2016.
- [34] Simone Milani. Compression of multiple user photo galleries. *Image and Vision Computing*, 53:68–75, 2016.
- [35] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531. IEEE, 2005.
- [36] Josef Sivic, Bryan C Russell, Alexei A Efros, Andrew Zisserman, and William T Freeman. Discovering objects and their location in images. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 370–377. IEEE, 2005.
- [37] Bryan C Russell, William T Freeman, Alexei A Efros, Josef Sivic, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1605–1614. IEEE, 2006.
- [38] Liangliang Cao and Li Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.
- [39] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International journal of computer vision*, 79(3):299–318, 2008.
- [40] Liangying Peng, Ling Chen, Xiaojie Wu, Haodong Guo, and Gencai Chen. Hierarchical complex activity representation and recognition using topic model and classifier level fusion. *IEEE Trans. Biomed. Engineering*, 64(6):1369–1379, 2017.
- [41] Jordan Boyd-Graber, Yuening Hu, David Mimno, et al. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296, 2017.
- [42] Shiran Dudy and Steven Bedrick. Ohsu@ mediaeval 2015: Adapting textual techniques to multimedia search. In *MediaEval*, 2015.
- [43] Kenneth Church and William Gale. Inverse document frequency (idf): A measure of deviations from poisson. In *Natural language processing using very large corpora*, pages 283–295. Springer, 1999.
- [44] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- [45] Shengsheng Qian, Tianzhu Zhang, Changsheng Xu, and Jie Shao. Multi-modal event topic model for social event analysis. *IEEE transactions on multimedia*, 18(2):233–246, 2016.

- [46] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42:177–196, 2001.
- [47] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001.
- [48] Thomas Hofmann. Probabilistic latent semantic indexing. In *ACM SIGIR Forum*, volume 51, pages 211–218. ACM, 2017.
- [49] Fan Yang, Wongun Choi, and Yuanqing Lin. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2129–2137, 2016.
- [50] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.
- [51] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [52] Dekang Lin et al. An information-theoretic definition of similarity. In *Icml*, volume 98, pages 296–304, 1998.
- [53] Lingling Meng, Runqing Huang, and Junzhong Gu. A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology*, 6(1):1–12, 2013.
- [54] George A Miller, Claudia Leacock, Randee Teng, and Ross T Bunker. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics, 1993.
- [55] Winthrop Nelson Francis, Henry Kučera, and Andrew W Mackie. *Frequency analysis of English usage: Lexicon and grammar*. Houghton Mifflin Harcourt (HMH), 1982.
- [56] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461. Springer, 2014.
- [57] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.
- [58] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.
- [59] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics, 2010.
- [60] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 262–272. Association for Computational Linguistics, 2011.
- [61] Susan T Dumais. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230, 2004.
- [62] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.
- [63] Gwonen Shieh. Choosing the best index for the average score intraclass correlation coefficient. *Behavior Research Methods*, 48(3):994–1003, Sep 2016.