A picture is worth a thousand words but how to organize thousands of pictures?

Stefan Lonn\textsuperscript{a}, Petia Radeva\textsuperscript{a,b}, Mariella Dimiccoli\textsuperscript{a,b,*}

\textsuperscript{a}University of Barcelona, Gran via de les Corts Catalanes 585, Barcelona 08007, Spain
\textsuperscript{b}Computer Vision Center, Edifici O Campus UAB, Bellaterra (Cerdanyola del Valles) 08193, Spain

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 10 persons. Experimental results demonstrate better user satisfaction with respect to state of the art solutions in terms of organization.

Keywords:
2000 MSC: 41A05, 41A10, 65D05, 65D17, hierarchical classification, probabilistic latent semantic analysis, convolutional neural networks

*Corresponding author: Tel.: +34667330940; fax: +34;
Email address: mariella.dimiccoli@ub.edu (Mariella Dimiccoli)
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 ([32, 34]). This has initiated extensive research on content-based image retrieval systems ([5, 6, 30, 8, 10, 2, 16]). 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.

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 smartphones’ 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 [28, 17, 37, 15, 33, 5, 6, 30, 27, 3, 10, 16, 2, 7], photo labelling [19, 14], and event recognition from single images shared online [1]. However, the approaches proposed so far are not directly applicable to smartphone pictures, since they lack of temporal structure and social network metadata, and presents a huge variability in terms of depicted objects, people 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. Softwares 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 former are estimated by means of an unsupervised approach, while the latter by a set of Convolutional Neural Networks (CNN). The important benefits of the proposed approach are: (i) a hierarchical organization in categories and subcategories that makes easier photo browsing, (ii) an unsupervised approach for category (topic) classification, and (iii) a very large number of categories for each topic that is of interest for people who have hobbies, or like to have pictures of a particular topic. As additional contribution, we make public a large subset of our testset in order to ease further investigation in the direction of personal smartphone photo organization. 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 \( \textcolor{red}{3} \) details the proposed approach. Sections \( \textcolor{red}{4.1} \) and \( \textcolor{red}{4.2} \) describe our experimental setting and discuss the experimental results, respectively. Finally, section \( \textcolor{red}{5} \) concludes the paper by highlighting the main contributions and outlining future work.

2. Related work

**Clustering and segmentation 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 (\[28, 17, 37, 15, 33, 27, 7\]). More specifically, time metadata, low-level information and, more recently, other picture metadata such as GPS have been used as features for these tasks. However, clustering and temporal segmentation 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 photo 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 (\[30, 3, 10, 2\]), or time metadata and GPS data (\[5, 6, 16\]) to automatically label events. For example, \[31\] 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 \[11\], 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 \cite{38} 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 fine predictions.

An original approach was taken by \cite{3} 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, \cite{2} 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, people and places.

*Event recognition from images.* Nowadays, a large number of photos captured by a smartphone or a digital camera is 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, \cite{1} 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 images.

Unlike this work that focuses on snapshots shared on-line which 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.

*Photo labeling.* More recent works have focused on indexing photos on the web shared on social networks such as Picasa, Flickr, Facebook and Instagram. These sharing photo communities generate vast amounts of metadata
as users interact with their images that have been exploited for multi-label annotation. [19] proposed a graphical model that explicitly accounts for the interdependencies between images sharing common properties that go beyond tags associated to images 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 [14] 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, [16] 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.

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 Photos\footnote{https://www.blog.google/products/photos/}, 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 which can be useful for pictures captured during a trip, but 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 PicJoy\footnote{https://www.picjoy.com/} 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 Photos\(^3\) classifies the users 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 2017\(^4\) 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 solution can be considered good only for photographers who are used to take care of their pictures timely and periodically, not by common smartphone users who typically have thousands of pictures automatically stored in the cloud and struggle to find a picture of interest or easily forget their pictures.

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.

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 ([12]). 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 \(i = 1, \ldots, M, j = 1, \ldots, N\) 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

\(^3\)https://itunes.apple.com/app/eden-photos-heavenly-simple/id1118761521  
\(^4\)http://www.toptenreviews.com/software/multimedia/best-photo-organizing-software/
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,

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

(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 $k \in \{1, \ldots, K\}$. The topic vectors are common to all documents in the corpus and the mixture vectors are specific to each document. For example, in Fig. 3 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 Maximization (EM) algorithm ([36]) 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, \( d_{\text{test}} \) and we would like to assign a topic to it. Given the distribution of words in the documents of the test set, say \( P(w|d_{\text{test}}) \), the document specific mixing coefficients, \( P(z|d_{\text{test}}) \) can be computed using the so called folding-in heuristic ([13]). When we have a new document \( d_{\text{test}} \), 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 \( P(z_k|d_{\text{test}}) \) are updated. In this way, we obtain the mixed coefficient \( P(z|d_{\text{test}}) \) 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:

\[
\arg\max_k P(z_k|d_{\text{test}}), \, k = 1, \ldots, K. \tag{2}
\]

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 ([35, 25]). 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 heuristics 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 of the actually 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 ([21]). 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. 4). We measure the semantic similarity between two words based on the shortest path in the hypernym taxonomy. Specifically, we used the Lin function (18, 20), which is an information content-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 \textit{mosquito} and \textit{bee} in WordNet is \textit{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,

\[
S(s_1, s_2) = \frac{2 \cdot IC(LCS(s_1, s_2))}{IC(s_1) + IC(s_2)}.
\]

where \( IC \) stands for Information Content that 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 \textit{SemCor} (22), a manually sense-tagged subset of the Brown Corpus (9).

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,

\[
\text{argmax}_k \sum_{i=1,\ldots,Q}^{j=1,2} S(s_i, s^k_j), k = 1, \ldots, K
\]

where \( s_i \) is the synset associated to a tag of the image and \( s^k_j \) 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 \textit{Street-view and Architecture} and \textit{Nature and Landscape}, we used the categories of the Places dataset (38), that contains 10,624,928 images from 434 categories. We ended up with 277 categories for \textit{Street-view and Architecture} and 88 categories for \textit{Nature and Landscape}.
respectively. For the Food and Drinks topic, we used all 101 categories of the Food101 dataset ([4]). For Sport and Adventure, we used the categories of the UCF Sports Action Dataset ([26]) more those relate to Sport of the WIDER dataset ([11]). For Interior and Object and Animal and Pets, we manually selected the appropriate categories from the ImageNet dataset ([8]) and the Places dataset ([38]). 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, screenshot, 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 ([3]): birthday, concert, exhibition, graduation, wedding. The total number of categories for each topic are detailed on Table 1.

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 2, 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 10 subjects, different from those who participated in the collection of the training set, for a total number of 8,749 images, with an
average of 875 pictures per user. Additionally, we downloaded the pictures of five Instagram’s vlogger, with an average of 683 images per user.

Table 1: Topic names and number of categories per topics

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

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.

*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 [24] and the UMass measure introduced by [23]. The UCI-score, $C_{UCI}$ uses as pairwise score function, the Pointwise Mutual Information (PMI) and is defined as follows:

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

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,
Figure 5: Visual interface used to show to the participants of the user study the results of two different systems. The top images correspond to the results obtained with our system and the bottom image to the results obtained with Eden photos.
and \( N \) is the total number of words. The smoothing count, \( \epsilon \) is added to avoid calculating the logarithm of zero. Note that \( C_{UCI} \) range is in the interval \( \{-1, 1\} \).

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 \( W = \langle w_1, ..., w_n \rangle \), it is defined as:

\[
C_{UMass} = \frac{2}{N(N-1)} \sum_{j=2}^{N} \sum_{i=1}^{j-1} \log \frac{P(w_j, w_i) + \epsilon}{P(w_i)}
\]

Note that it 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:

\[
AvgNPMI = \frac{1}{Q(Q-1)} \sum_{j=2}^{N} \sum_{i=1}^{j-1} \log \frac{P(w_j, w_k)}{P(w_i, w_k)}
\]

where \( k \) indicates the k-th topic.

*Assessing the proposed organization through an user study.* The proposed approach has been evaluated through user studies, since ultimately the impact of the automatic organization depends on its value to the user. As subjects, we recruit both the 10 owners of the photo collections as well as 10 subjects uninvolved 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.

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, we showed to the users two screens at the same time: in one the organization of our method and in the other - a competitive method. To avoid bias judgment due to different visualization, we mimicked the visual interface of Google Photos and presented the results
Table 3: Architecture and initial weights used for the training of 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                       |

using the same interface (see Fig. 5). We asked to the participants two questions: The first question that evaluates aspect a), was: *Which kind of organization do you prefer and why?* 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 *Nature and Landscape*, 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 technology[^1]. 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

[^1]: [http://www.imagga.com/solutions/auto-tagging.html](http://www.imagga.com/solutions/auto-tagging.html)
experiments), the picture is assigned to the \textit{Null} topic.

As it can be appreciated in Table 1, 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 \textit{Sunrises and Sunset, Beaches and Seaside} and \textit{Flowers} can be considered as categories of \textit{Nature and Landscape} instead of being a topic itself. Similarly, we treated \textit{Painting and Art} a category of \textit{Text and Visual} and \textit{Cars and Vehicles} as a subcategory of \textit{City and Architecture}.

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 studies.

The images were then 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 3. 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.

- **Nature and Panorama**: europe, rocks, shoreline, barrier, surf, seaside, asia, boundary, sunrise, hill, sunshine, seashore, ship, vessel,
Table 5: 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 10 users. Numbers indicate percentage of responses for each question.

|                      | Photos owners |                  |                  |                  |                  |
|----------------------|---------------|------------------|------------------|------------------|------------------|
|                      | Much better   | Better | Similar | Worse | Much worse |
| Organization         | 60%           | 40%    | 0%     | 0%    | 0%           |
| Accuracy             | 0%            | 30%    | 30%    | 40%    | 0%           |
| External evaluators  | 70%           | 30%    | 0%     | 0%    | 0%           |
| Accuracy             | 6.67%         | 33.33% | 46.67% | 13.33% | 0%           |

|                      | Photos owners |                  |                  |                  |                  |
|----------------------|---------------|------------------|------------------|------------------|------------------|
|                      | Much better   | Better | Similar | Worse | Much worse |
| Organization         | 60%           | 40%    | 0%     | 0%    | 0%           |
| Accuracy             | 0%            | 40%    | 40%    | 20%    | 0%           |
| External evaluators  | 40%           | 50%    | 10%    | 0%    | 0%           |
| Accuracy             | 13.33%        | 40.00% | 36.67% | 10.00% | 0%           |

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, res-
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 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 | Organization       | 60%    | 40%     | 0%    | 0%         |
| Photos | Accuracy          | 20%    | 10%     | 30%   | 40%        | 0%         |

|        | Much better | Better | Similar | Worse | Much worse |
|--------|-------------|--------|---------|-------|------------|
| Eden   | Organization       | 60%    | 40%     | 0%    | 0%         |
| Photos | Accuracy          | 0%     | 30%     | 20%   | 50%        | 0%         |

Figure 7: Example of images correctly classified by our system

grotto, exhibition, escalator, indoor, screenshot, wave, paella, altar, crosswalk, brittany spaniel, hare

Fig. 6 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. In Table 4, we report the values of the topic coherence measures described in section 4.1.2, obtained by using the Movie corpus, a Wikipedia subset, as external corpus in the gensim Python library.
It can be observed that the three topic measures show consistent results. The topic *Sport and Exercise* is the less coherent whereas *Nature and Panorame* and *Art and visual* are the most coherent.

### 4.2.2. User study results

We recruited 20 persons for the user study. In average, each photo collection has been evaluated by three different participants. Half of them were not involved with the data collection, and three of them were ‘computer illiterate’. 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, Travel, Panorama, Selfies, Food, Documents, Dogs, Sport (described with the favorite one such as Skatering).

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 1).

### Table 7: Event classes used by state of the art algorithms

| PEC | Holiday 1 | Holiday 2 | SocEID |
|-----|-----------|-----------|--------|
| 1   | Birthday  | Holiday    | 40     |
| 2   | Mardi gras| Beach fun  | 20     |
| 3   | Christmas | Christmas  | 40     |
| 4   | Children  | Graduation | 10     |
| 5   | Graduation| Halloween  | 10     |
| 6   | Hollyday  | Graduations| 10     |
| 7   | Christmas | Graduation | 10     |
| 8   | Christmas | Graduation | 10     |
| 9   | Thanksgiving| Graduation | 10     |
| 10  | Thanksgiving| Graduation | 10     |
| 11  | Thanksgiving| Graduation | 10     |
| 12  | Graduation | Graduation | 10     |
| 13  | Graduation | Graduation | 10     |
| 14  | Graduation | Graduation | 10     |
| 15  | Graduation | Graduation | 10     |
| 16  | Graduation | Graduation | 10     |
| 17  | Graduation | Graduation | 10     |
| 18  | Graduation | Graduation | 10     |
| 19  | Graduation | Graduation | 10     |
| 20  | Graduation | Graduation | 10     |
| 21  | Graduation | Graduation | 10     |
| 22  | Graduation | Graduation | 10     |
| 23  | Graduation | Graduation | 10     |
| 24  | Graduation | Graduation | 10     |
| 25  | Graduation | Graduation | 10     |
| 26  | Graduation | Graduation | 10     |
| 27  | Graduation | Graduation | 10     |
| 28  | Graduation | Graduation | 10     |
| 29  | Graduation | Graduation | 10     |

**Rare Event Dataset**

| Eden | WIDER | UIUC Sports |
|------|-------|-------------|
| 1    | Parade| Soldier drilling |
| 2    | Handshaking| Spa |
| 3    | Hurricane Katrina| Demonstration |
| 4    | Riot| Students Schoolkids |
| 5    | Nepal earthquake| Dancing |
| 6    | Car accident| Lawyer- Waitress |
| 7    | Obama wins elections| Funeral |
| 8    | Columbia space shuttle disaster| Cheering |
| 9    | Election Trump| Press conference |
| 10   | Attack 9/11| Ancestry |
| 11   | Trump wins election| Voting |
| 12   | Russia wins election| Resigning |
| 13   | Putin wins election| Resigning |
| 14   | Putin wins election| Resigning |
| 15   | Putin wins election| Resigning |
| 16   | Putin wins election| Resigning |
| 17   | Putin wins election| Resigning |
| 18   | Putin wins election| Resigning |
| 19   | Putin wins election| Resigning |
| 20   | Putin wins election| Resigning |
| 21   | Putin wins election| Resigning |
| 22   | Putin wins election| Resigning |
| 23   | Putin wins election| Resigning |
| 24   | Putin wins election| Resigning |
| 25   | Putin wins election| Resigning |
| 26   | Putin wins election| Resigning |
| 27   | Putin wins election| Resigning |
| 28   | Putin wins election| Resigning |
| 29   | Putin wins election| Resigning |

---

22
As it can be appreciated on Table 5 and Table 6, 100% of the participants considered the proposed organization "much better" or "better" than the organization proposed by both Google Photos and Eden Photos. The photo owners judged the accuracy of our system slightly worse than those of Google Photos, whereas participants who saw the pictures for the first time, judged the accuracy slightly better. With Eden Photos, we observed a different trend: the photo owners judged our system slightly more accurate than external participants, but both considered our system more accurate than Eden. It is worth to take into account that only the Eden Photos app classifies all pictures, whereas the Google Photos app classifies in average 75.4% of the pictures, whereas our system classifies 89% of the pictures.

All systems, included ours, performed slightly worse on the images of the vloggers, since these have been filtered before being posted on Instagram. On these pictures, the accuracy of our system was judged slightly better (see Table 6).

It is very important to remark that Google Photos always classifies the images into a relatively small set of categories, in average 22 over the 10 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. Furthermore, the only category related to people that was found in the full testing set was selfies. 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 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.

4.2.3. Qualitative results and comparisons

Fig. 7 and 8 show examples of pictures correctly and incorrectly classified by our system, respectively. In particular, in Fig. 7 it is possible to appreciate the level of detail that can be achieved by our system. On the first row of Fig. 8 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 crow
are easily wrongly assigned to Sports and Adventure.

Existing algorithms for event recognition from personal photo collections have focused on the detection of a limited set of social events (see Table 7). 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 7) have been used in the literature ([5, 6, 31, 29, 31, 10, 1]). 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.

Our system is not able to recognize rare events such the ones of the Rare Event Dataset (see Table 7). However, as demonstrated by the literature on event recognition, the use of time metadata and GPS information would be extremely useful in detecting events. In particular, in a smartphone, it would be useful to use information coming from the Google Calendar to recognize events that happen only in a specific day of the year. 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 such as Eden Photos and Google Photos. The proposed approach could be easily integrated in a retrieval system that relies on both semantic tags and time metadata to retrieve all images corresponding to the user query. With the goal of encouraging research on smartphone picture organization, we make available the test set and the semantic tags of the training set.

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. Finally, we gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used in our research. This work was partially funded by TIN2015-66951-C2-1R, SGR-1219, SGR-1742, Grant 20141510 (Marató TV3), CERCA Programme / Generalitat de Catalunya and ICREA Academia grant.

References

[1] Ahsan, U., Sun, C., Hays, J., Essa, I., 2017. Complex event recognition from images with few training examples, in: Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on, IEEE. pp. 669–678.

[2] Bacha, S., Allili, M.S., Benblidia, N., 2016. Event recognition in photo albums using probabilistic graphical models and feature relevance. Journal of Visual Communication and Image Representation 40, 546–558.

[3] Bossard, L., Guillaumin, M., Van Gool, L., 2013. Event recognition in photo collections with a stopwatch hmm, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1193–1200.

[4] Bossard, L., Guillaumin, M., Van Gool, L., 2014. Food-101–mining discriminative components with random forests, in: European Conference on Computer Vision, Springer. pp. 446–461.

[5] Cao, L., Luo, J., Huang, T.S., 2008. Annotating photo collections by label propagation according to multiple similarity cues, in: Proceedings of the 16th ACM international conference on Multimedia, ACM. pp. 121–130.

[6] Cao, L., Luo, J., Kautz, H., Huang, T.S., 2009. Image annotation within the context of personal photo collections using hierarchical event and scene models. IEEE Transactions on Multimedia 11, 208–219.

[7] Datia, N., Pires, J.M., Correia, N., 2017. Time and space for segmenting personal photo sets. Multimedia Tools and Applications 76, 7141–7173.

[8] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE. pp. 248–255.
[9] Francis, W., Kucera, H., 1982. Frequency analysis of english usage.

[10] Gu, J.M., Wu, Y.L., Hung, W.C., Tang, C.Y., 2013. Personal photo organization using event annotation, in: Information, Communications and Signal Processing (ICICS) 2013 9th International Conference on, IEEE. pp. 1–4.

[11] Guo, C., Tian, X., 2015. Event recognition in personal photo collections using hierarchical model and multiple features, in: Multimedia Signal Processing (MMSP), 2015 IEEE 17th International Workshop on, IEEE. pp. 1–6.

[12] Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. Machine learning 42, 177–196.

[13] Hofmann, T., 2017. Probabilistic latent semantic indexing, in: ACM SIGIR Forum, ACM. pp. 211–218.

[14] Johnson, J., Ballan, L., Fei-Fei, L., 2015. Love the neighbors: Image annotation by exploiting image metadata, in: Proceedings of the IEEE international conference on computer vision, pp. 4624–4632.

[15] Kang, H., Bederson, B.B., Suh, B., 2007. Capture, annotate, browse, find, share: Novel interfaces for personal photo management. International Journal of Human- Computer Interaction 23, 315–337.

[16] Kumar, N., Seitz, S., 2014. Photo recall: Using the internet to label your photos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 771–778.

[17] Latif, K., Mustofa, K., Tjoa, A.M., 2006. An approach for a personal information management system for photos of a lifetime by exploiting semantics, in: DEXA, Springer. pp. 467–477.

[18] Lin, D., et al., 1998. An information-theoretic definition of similarity., in: Icml, pp. 296–304.

[19] McAuley, J., Leskovec, J., 2012. Image labeling on a network: using social-network metadata for image classification. Computer Vision–ECCV 2012 , 828–841.
[20] Meng, L., Huang, R., Gu, J., 2013. A review of semantic similarity measures in wordnet. International Journal of Hybrid Information Technology 6, 1–12.

[21] Miller, G.A., 1995. Wordnet: a lexical database for english. Communications of the ACM 38, 39–41.

[22] Miller, G.A., Leacock, C., Tengi, R., Bunker, R.T., 1993. A semantic concordance, in: Proceedings of the workshop on Human Language Technology, Association for Computational Linguistics. pp. 303–308.

[23] Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A., 2011. Optimizing semantic coherence in topic models, in: Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics. pp. 262–272.

[24] Newman, D., Lau, J.H., Grieser, K., Baldwin, T., 2010. Automatic evaluation of topic coherence, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics. pp. 100–108.

[25] Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788.

[26] Rodriguez, M.D., Ahmed, J., Shah, M., 2008. 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, IEEE. pp. 1–8.

[27] Sun, F., Li, H., Wang, X., 2013. Photo 4w: Mobile photo management on what, where, who and when. Neurocomputing 119, 59–64.

[28] Sun, Y., Zhang, H., Zhang, L., Li, M., 2002. Myphotos: a system for home photo management and processing, in: Proceedings of the tenth ACM international conference on Multimedia, ACM. pp. 81–82.

[29] Tang, F., Tretter, D.R., Willis, C., 2011. Event classification for personal photo collections, in: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, IEEE. pp. 877–880.
[30] Tsai, S.F., Cao, L., Tang, F., Huang, T.S., 2011a. Compositional object pattern: a new model for album event recognition, in: Proceedings of the 19th ACM international conference on Multimedia, ACM. pp. 1361–1364.

[31] Tsai, S.F., Huang, T.S., Tang, F., 2011b. Album-based object-centric event recognition, in: Multimedia and Expo (ICME), 2011 IEEE International Conference on, IEEE. pp. 1–6.

[32] Van House, N.A., 2009. Collocated photo sharing, story-telling, and the performance of self. International Journal of Human-Computer Studies 67, 1073–1086.

[33] Viana, W., Bringel Filho, J., Gensel, J., Villanova-Oliver, M., Martin, H., 2008. Photomap: from location and time to context-aware photo annotations. Journal of Location Based Services 2, 211–235.

[34] Whittaker, S., Bergman, O., Clough, P., 2010. Easy on that trigger dad: a study of long term family photo retrieval. Personal and Ubiquitous Computing 14, 31–43.

[35] Yang, F., Choi, W., Lin, Y., 2016. 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, pp. 2129–2137.

[36] Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. IEEE transactions on medical imaging 20, 45–57.

[37] Zhao, M., Teo, Y.W., Liu, S., Chua, T.S., Jain, R., 2006. Automatic person annotation of family photo album, in: International Conference on Image and Video Retrieval, Springer. pp. 163–172.

[38] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2017. Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence.
Supplementary Material

Sport and Adventure

1. Lacrosse
2. Curling bonspiel
3. Archery
4. Skiing
5. Kayaking
6. Bmx
7. Tennis
8. Gymnastics
9. Rowing
10. Polo
11. Table tennis
12. Volleyball
13. Sumo
14. Rugby
15. Bowling
16. Handball
17. Snowboard
18. Basketball
19. Surfing
20. Rock climbing
21. Badminton
22. Boxing
23. Wrestling
24. Parachute sport
25. Equitation
26. Fishing lake
27. Mountain biking
28. Water polo
29. Swimming
30. Golf
31. Weightlifting
32. Skateboarding
33. Diving
34. Cycling
35. Sailing
36. Hockey
37. Baseball
38. Running sport
39. Soccer
40. Croquet

Text and Visual
1. Map
2. Screenshot
3. Magazines
4. Drawing
5. Sign
6. Tattoo
7. Poster
8. Graffiti
9. Painting
10. Receipt
11. Writing