Cluster-Based Landmark and Event Detection for Tagged Photo Collections

Symeon Papadopoulos, Christos Zigkolis, and Yiannis Kompatsiaris
Centre for Research and Technology Hellas

Athena Vakali
Aristotle University

An image analysis scheme can automate the detection of landmarks and events in large image collections, significantly improving the content-consumption experience.

The rising popularity of photo-sharing applications on the Web has led to the generation of huge amounts of personal image collections. Browsing through image collections of such magnitude is currently supported by the use of tags. However, tags suffer from several limitations—such as polysemy, lack of uniformity, and spam—thus not presenting an adequate solution to the problem of content organization. Therefore, automated content-organization methods are of particular importance to improve the content-consumption experience. Because it's common for users to associate their photo-captured experiences with some landmarks—for example, a tourist site or an event, such as a music concert or a gathering with friends—we can view landmarks and events as natural units of organization for large image collections. It's for this reason that automating the process of detecting such concepts in large image sets can enhance the experience of accessing massive amounts of pictorial content.

In this article, we present a novel scheme for automatically detecting landmarks and events in tagged image collections. Our proposal is based on the simple yet elegant concept of image similarity graphs as a means of combining multiple notions of similarity between images in a photo collection; in our case, we use visual and tag similarity. We perform clustering on such image similarity graphs by means of community detection,[1] a process that identifies on the graph groups of nodes that are more densely connected to each other than to the rest of the network. In contrast to conventional clustering schemes such as *k*-means or hierarchical agglomerative clustering, community detection is computationally more efficient and doesn't require the number of clusters to be provided as input. Subsequently, we classify the resulting image clusters as landmarks or events by use of features related to the temporal, social, and tag characteristics of image clusters. In the case of landmarks, we also conduct a cluster-merging step on the basis of spatial proximity to enrich our landmark model.

Landmark- and event-detection framework

Image groups are extracted from the original tagged-image collection by means of a graph-based image-clustering algorithm that operates on a hybrid image-similarity graph, including visual and tag similarities between images. Subsequently, the image clusters found by this algorithm are classified as either landmarks or events. Landmark clusters are merged on the basis of their spatial proximity and labeled by use of some additional tag processing. Figure 1 depicts an overview of the framework.

Hybrid image clustering

The proposed image-clustering framework relies on the creation of two image graphs representing two kinds of similarity between images, with the similarity being based on their visual features and their tags. Subsequently, community detection, consisting of an efficient graph-based clustering scheme, is applied on the union of these two graphs to identify sets of nodes (that is, the image clusters) that are more densely connected to each other than to the rest of the network.

For visual-similarity graph creation, we compute the scale invariant feature transform (SIFT)

1070-986X/11/$26.00 © 2011 IEEE

Published by the IEEE Computer Society

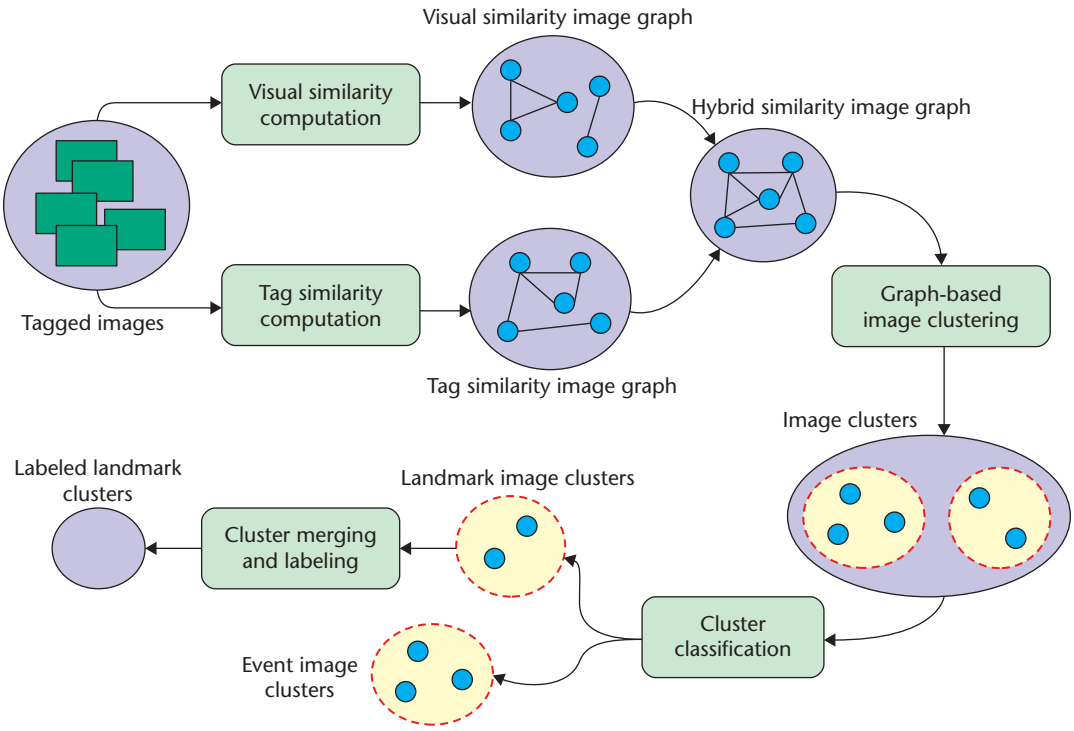

Figure 1. Overview of proposed framework for detecting landmarks and events within tagged-image collections.

for every image. The original 128-dimensional SIFT descriptors were introduced by Lowe.[2] In our experiment, we accomplish the extraction using the software implementation of Van de Sande, Gevers, and Snoek.[3] We use a bag-of-words model with a vocabulary of 500 words and perform the clustering process using the *k*-means algorithm. The assignment of visual words to image features is performed using the code word uncertainty model.[3] Once the visual feature vectors are extracted, pairwise similarities between images (using some similarity function, such as cosine similarity or inverse Minkowsky distance) are computed and the top-K ($K = 20$) most similar images for each image are inserted as its neighbors on the graph. Subsequently, we specify a visual-similarity threshold as the median of all similarities in the graph and filter out all edges falling below this threshold.

We base the creation of the tag similarity graph on the co-occurrences of tags in the image contexts. We process the image-tag associations to build an inverted index, which maintains for each tag a list of annotated images. Each possible pair of images in this list leads to the creation of an undirected edge between these two images on the image graph. The edge is weighted by the number of times these two images are found together in a tag list. Tags associated with long image lists (that is, used frequently to tag images) are excluded from the process of establishing links between images on the graph. In that way, we avoid the insertion of spurious edges in the tag-similarity graph, which would only indicate an obvious association between two images through some generic tag. Moreover, this process leads to considerable computational gains because the number of all possible pairs in a list of length n (that we avoid considering) is $n \cdot (n - 1)/2 \propto n^2$.

After the creation of the tag-similarity graph, we filter out edges with co-occurrence frequency below an empirically selected threshold. Such a filtering step is designed to remove associations among images that are not common and in addition makes the problem of graph clustering easier from a computational perspective because the resulting graph is sparser.

An alternative, more precise, but also more computationally expensive approach for deriving the tag-similarity graph is by use of tag-based features. A typical approach often used in related problems[4] is to consider the image-tag occurrence matrix formed by the given collection of tagged images. In this way, images are represented as vectors in the space of unique tags. Due to the high dimensionality

of the raw vector space, it's customary to apply an appropriate dimensionality reduction technique, such as latent semantic indexing (LSI), which projects the raw tag space on a space of much fewer latent dimensions. Then, pairwise similarities between images are computed in the latent space and the top-K most similar images for each image are inserted as its neighbors on the graph. Finally, thresholding similar to the one used for the visual graph creation is applied. Such an approach results in a more accurate and noise-resilient depiction of tag-based image similarities, but it comes at a substantial computational overhead.

Once the visual- and tag-similarity graphs are created, they are merged into a hybrid image graph comprising the union of their nodes and the union of their edges. On the basis of these three image similarity graphs, we perform graph-based image clustering by use of community detection, that is, by identifying regions on the graph that are more densely connected to each other than to the rest of the network.[1] We have experimented with the structural clustering algorithm for networks (SCAN) approach,[5] which is based on the concept of structural similarity between nodes. The structural similarity between nodes u and w is defined as

$$\sigma(u, w) = \frac{|\Gamma(u) \cap \Gamma(w)|}{\sqrt{|\Gamma(u)| \cdot |\Gamma(w)|}}$$

where $\Gamma(u)$ is the structure of node u, that is, the set comprising the node's neighbors and the node itself as elements. Communities (clusters) are then defined as groups of μ nodes that have a structural similarity value of at least ε between each other ((μ, ε)-cores[5]). The rest of the nodes on the graph are not assigned to any cluster, leading in that way to clusters of high coherence, meaning they contain members that are highly related to each other.

Compared to conventional schemes used previously for the problem of image clustering—such as k-means and hierarchical agglomerative clustering—community detection offers improved computational efficiency. With this approach, there's no need to know the number of clusters to be extracted. In addition, this approach offers the possibility of tuning cluster coherence by means of the two parameters (μ and ε). Increasing μ results in fewer and larger communities while increasing ε makes the clustering scheme more selective and leaves more images unassigned to any cluster. The option of keeping unrelated images out of the produced cluster structure is particularly important in the context of social-content sources because such content often entails significant amounts of noise.

In terms of computational demands, the method used in this study scales with $O(k_m \cdot m)$, where k_m stands for the average degree of the graph and m is the number of graph edges. In practice, because k_m is upper-bounded due to the nature of the graph-construction process (retaining the top-K most similar images per image), the complexity of the clustering method is $O(m)$. This method offers a much lower complexity than conventional clustering schemes, such as $O(I \cdot C \cdot n \cdot D)$ for k-means and $O(n^2 \cdot \log n)$ for hierarchical agglomerative clustering, where n stands for the number of objects, I the number of iterations, C the number of clusters, and D the number of dimensions of the feature vectors. In terms of memory requirements, the employed community-detection technique is also advantageous because it needs approximately $2 \cdot (n + m)$ memory elements, while k-means needs $(n + C) \cdot D$ and hierarchical agglomerative clustering needs n^2 memory elements.

Cluster classification

Once clusters of images have been extracted by the process described previously, each cluster is classified as either a landmark or event. To proceed with this classification, we employ several standard classification algorithms (namely kNN and support vector machine, or SVM), which use four features for each cluster. Two of these features, which constitute our baseline, were introduced in Quack, Leibe, and Van Gool: the duration of the cluster in days (computed by subtracting the timestamp of the earliest image of the cluster from the one of the most recent), and the ratio of the number of unique image creators over the number of images in the cluster.[6] We denote the first feature as $f_1 = |D|$, where $|D|$ stands for the number of days spanned by the image cluster and the second as $f_2 = |U|/N$, where $|U|$ is the number of unique users contributing images to this cluster and N is the number of images in the cluster.

Quack, Leibe, and Van Gool show these features to be effective in distinguishing between landmark and event image clusters. The main motivation behind their use is that landmarks are photographed by many people and throughout long periods. In contrast, events are usually characterized by a short duration (up to few days) and covered by few people. In practice, however, there are numerous cases that these features are not discriminative enough for the purpose of landmark and event classification. For instance, during our experimentation we ran into several clusters of images, which, despite the fact that the clusters consisted solely of images by a single user, depicted landmarks. Had we relied exclusively on these two features, such cases would have definitely been misclassified. A similar situation arises in cases where a cluster comprises multiple similar events (for example, weddings) that span a long period. The features of such a cluster would result in it erroneously being classified as a landmark in the aforementioned feature space.

To address this limitation of the cluster feature space, we propose the use of two additional features that are based on the cluster images' tags. Because we have a set of training clusters at our disposal, labeled as either landmarks or events, we are able to create two tag profiles corresponding to the two cluster classes (landmark and event) in the form of tag frequency vectors. After deriving such tag vectors, we can identify the shared tags and then remove them from both cluster classes. In that way, we end up with a tag vector consisting of landmark-only tags and one consisting of event-only tags. For instance, landmark-only tags for an image collection focused on Barcelona include the tags ''gaudi,'' ''architecture,'' ''buildings,'' ''railway,'' ''park,'' and so on; while event-only tags include ''concert,'' ''music,'' ''racing,'' ''live,'' and so on. Then, for each cluster, we can count the number of times that a tag from its images belongs to one set or another. These two counts constitute the two additional cluster features.

Landmark cluster merging and labeling

After the cluster-classification step, we apply an additional cluster-processing step on the image clusters that depict landmarks. The need for such a step stems from our observation that many of the landmark clusters refer to the same object. To maximize the utility of our image-organization framework, we would like all these clusters to be grouped together and be labeled with a meaningful name.

For the cluster-merging step, we make use of the geolocation information that is frequently available in tagged images. Based on this geolocation data, we derive for each image cluster two geographical features: geographical center by averaging over the geocoordinates of its geotagged images, and mean distance between cluster images. Based on these features, we create a spatial-proximity graph comprising image clusters as nodes and their pairwise distances as edges. We filter out edges exceeding a distance threshold of 300 meters and nodes, of which the mean distance between their images is higher than 300 meters. Once such a graph is formed, we apply our community-detection scheme and we merge image clusters belonging to the same community into metaclusters.

To assign a meaningful label to each metacluster, we aggregate their tags and rank them by frequency of appearance in the given metacluster. Subsequently, we discard tags that appear in more than two metaclusters as we consider these to be generic. Finally, we select the first five tags as the label of each metacluster. For summarizing each metacluster, we also randomly select one image from each of its containing clusters.

Evaluation

Our goal for the evaluation is to demonstrate that the resulting image clusters are suitable for the task of landmark and event detection and that detected clusters are classified with satisfactory precision to these classes. We conducted two comparisons: image clusters derived from the proposed graph-based method were compared to the ones derived from a baseline clustering scheme, namely *k*-means, and the proposed cluster feature space for landmark and event detection was compared to the baseline.[6] In all cases, the proposed techniques were able to produce superior results. The evaluation demonstrates that the detected landmarks can be automatically labeled and located with satisfactory precision and that the detected events span a wide range of types of both public and personal interest.

We conducted our experiments on a set of 207,750 images collected by querying Flickr with a geoquery centered on Barcelona. These

images were uploaded by 7,768 users. We first processed the image tags by filtering long and short tags, tags that consisted of both numeric characters and letters, as well as tags from a manually created blacklist. We also merged tags that were lexically similar to each other as expressed by the Levenshtein distance. Doing so resulted in a total of 33,959 unique tags, and 173,825 images tagged with at least one of them. Subsequently, we formed the tag-image list index and removed tags used in more than 350 images. Examples of such tags that can be considered uninformative for the particular data set include ''Barcelona,'' ''Spain,'' and ''Catalunya.'' This step reduced the unique tags to 33,367 and the images tagged with at least one of them to 120,742. Furthermore, out of the original set of 207,750 images, there were 195,308 geotagged images.

As a first step, we conducted a comparison between the cluster quality derived from the community-detection scheme and the conventional *k*-means clustering that is often used other work.[7,8] With the data set images as a starting point, we first formed the image-similarity graphs according to the process described previously. We created four image-similarity graphs for representing the visual similarity (VIS); the two variants of tag similarity, namely co-occurrence based (TAG-C) and LSI-based (TAG-LSI); and hybrid similarity (HYB) between images. The VIS and TAG-LSI graphs were built by use of the inverse city-block distance (that is, Minkowsky distance with $p = 1$). We built the HYB graph by considering the union of VIS and TAG-C graphs. Their sizes were respectively (137K, 2M), (83K, 3.6M), (92K, 1.3M), and (162K, 5.5M), where K stands for thousand, M for million, (n, m) for a graph of n nodes and m edges.

Subsequently, we applied the community-detection algorithm on each of the three graphs (for clustering the graphs, we set $\mu = 4$ and $\varepsilon = 0.6$.). To have a direct comparison with *k*-means, we also clustered separately the images contained in the VIS graph and the ones contained in the TAG-LSI graph using the SIFT-based visual and latent tag features, respectively. We set the number of clusters for *k*-means to be M, $2 \cdot M$, and $3 \cdot M$, where M is the number of communities produced by the respective community-detection method. Because there is no straightforward means of combining the visual and tag-based features by use of *k*-means, we compare separately the clustering performance for each kind of feature. Thus, we ended up with two different groups of clustering outputs: the ones based on visual features, namely SCAN_{VIS} and $\text{KM}_{\text{VIS},x\text{M}}$ (where $x = 1, 2, 3$); and the ones based on tag features, namely $\text{SCAN}_{\text{TAG-C}}$, $\text{SCAN}_{\text{TAG-LSI}}$, and $\text{KM}_{\text{TAG},x\text{M}}$, (where $x = 1, 2$).

For each clustering output, we derive two measures of quality: geospatial cluster coherence (GCC) and subjective cluster quality (SCQ). GCC is computed by use of the geotagging information available for most images in our collection. More specifically, for each cluster we compute the average geodesic distance between the cluster members and the cluster geographical center. Then, GCC is expressed as the mean and standard deviation of this quantity across all clusters. GCC constitutes an objective measure that can be automatically computed for the whole data set and thus makes possible a large-scale evaluation of the clustering quality. In contrast, SCQ was evaluated by human inspection of the clustering results on a set of 33 visual and 40 tag-based and randomly selected clusters.

More specifically, for each type of feature (visual and tag) a number of clusters (33 visual and 40 tag-based) were randomly selected from the results of SCAN and the corresponding clusters (those with the highest overlap in terms of contained members) from the KM results were also selected. We showed the different clusters to the users, asking them to mark the images in each cluster that weren't relevant to the main object or entity depicted by the cluster. The evaluators were not aware of the method that produced the clusters. Once the irrelevant images were marked for each cluster, it was possible to compute the typical information-retrieval performance measures (precision, recall, and *F*-measure) for each of the competing methods (recall was computed with reference to the total number of relevant images across the different clusterings). In addition, because each cluster was subjected to evaluation by two independent evaluators, it was possible to compute a κ-statistic for each method (interannotator agreement). Table 1 presents the collected results.

Observation of the tabulated results reveals that image clusters produced by the employed community-detection method (SCAN) are

Table 1. Cluster quality comparison between SCAN and k*-means approaches. The performance is evaluated separately on visual and tag-based features and for multiple values of* k*. We could not include* k*-means with* K = 3M *in the tag cluster comparison because the large number of* K *led to an estimated execution time of over a week.*

Cluster type	Clustering method (number of clusters)	Geospatial cluster coherence (*m* stands for meters) μ_d (m)	σ_d (m)	Subjective cluster quality *P*	*R*	*F*	κ
Visual	$SCAN_{VIS}$ (560)	357.1	1185.7	1.000	0.110	0.199	1.000
	$KM_{VIS,1M}$ (560)	2470.0	1734.4	0.806	0.324	0.462	0.226
	$KM_{VIS,2M}$ (1,120)	2249.7	1893.7	0.899	0.294	0.443	0.544
	$KM_{VIS,3M}$ (1,680)	2183.1	2027.4	0.929	0.271	0.420	0.719
Tag	$SCAN_{TAG-C}$ (1,774)	767.4	1712.0	0.898	0.253	0.394	0.642
	$SCAN_{TAG-LSI}$ (4,027)	456.3	1151.1	0.950	0.182	0.306	0.820
	$KM_{TAG,1M}$ (4,027)	766.8	1762.7	0.848	0.307	0.451	0.564
	$KM_{TAG,2M}$ (8,054)	563.2	1528.7	0.903	0.258	0.401	0.707

more precise than the ones produced by *k*-means clustering. In terms of the GCC measure, the SCAN-produced clusters are clearly superior to the *k*-means ones, which indicates better geographical focus and thus better correspondence to landmarks and events (which are usually highly localized). The difference in GCC is especially pronounced for visual clusters. The actual GCC performance of *k*-means clustering is worse than presented in Table 1 because in the computation of μ_d we also included one-member clusters (that obviously have $\mu_d = 0$), which constitute a large portion of the *k*-means results and thus reduce the average μ_d.

In terms of SCQ (subjective evaluation), once more the SCAN clusters appear to be of higher precision than the competing *k*-means ones. Clusters produced by *k*-means present considerably higher recall. However, the low κ-statistic values that characterize *k*-means clusters indicate that users don't agree on whether the images included in the clusters are related to each other. In contrast, there is a high interannotator agreement with respect to SCAN clusters. That means that these clusters are easy to interpret by users. Another interesting observation pertains to the cluster quality derived by each variant of tag-based graph similarity, namely plain co-occurrence and LSI. There is a clear cluster quality advantage in favor of LSI, which comes at a significant computational cost during the graph construction step.

We also conducted a similar study where we compared the quality of clusters derived by clustering the VIS, TAG-C, and HYB image similarity graphs. We found that the best information-retrieval performance is achieved by use of the hybrid similarity graph. More specifically, the *F*-measure of the HYB image clusters was 28.5 percent higher than the one of VIS clusters and 19.8 percent higher than the one of TAG-C clusters. The interannotator agreement for these results was substantial, because in all cases the obtained κ-statistic values were above 0.6. These results are consistent with our previous study.[9]

We also inspected the distribution of the different clusters (visual, tag, and hybrid) in the feature space of Quack, Leibe, and Van Gool. Figure 2 (next page) presents a comparison among the three clusterings, revealing conspicuous differences. One could postulate that visual clusters largely correspond to landmarks, while tag clusters mainly correspond to events. It's the hybrid clusters that span the whole feature space and thus correspond to both landmarks and events. Thus, we use hybrid clusters for the rest of the evaluation study.

To further quantify the purity of the resulting clusters in terms of correspondence to landmarks and events, we asked two users to look at the images of 60 hybrid clusters and provide a characterization of landmark or event at the image level—that is, to decide whether each image (seen in the context of the rest of the images belonging to the same cluster) depicted a landmark or an event. The annotation set of 60 clusters consisted of 1,127 images. For each cluster, we computed the percentage of images that were annotated as landmarks and

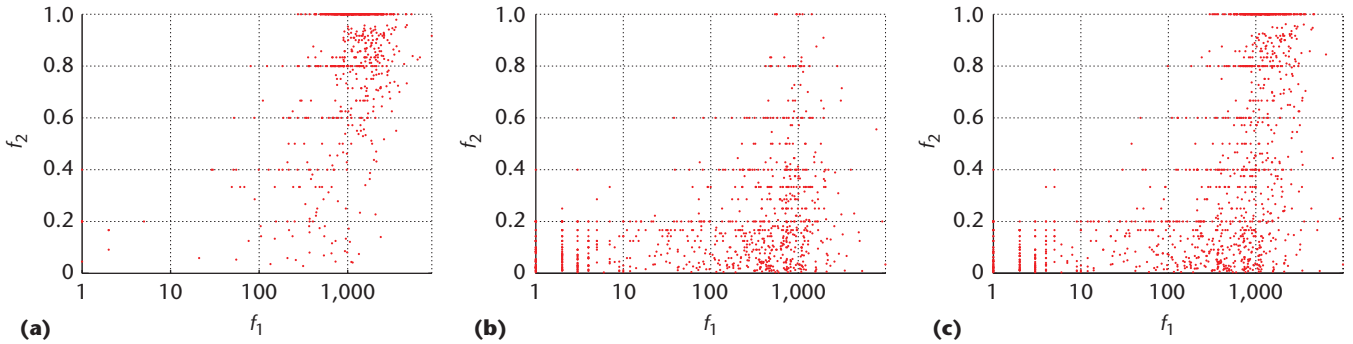

Figure 2. Comparison between image clusters derived from different similarity graphs: (a) $SCAN_{VIS}$, (b) $SCAN_{TAG\text{-}C}$, and $SCAN_{HYB}$.

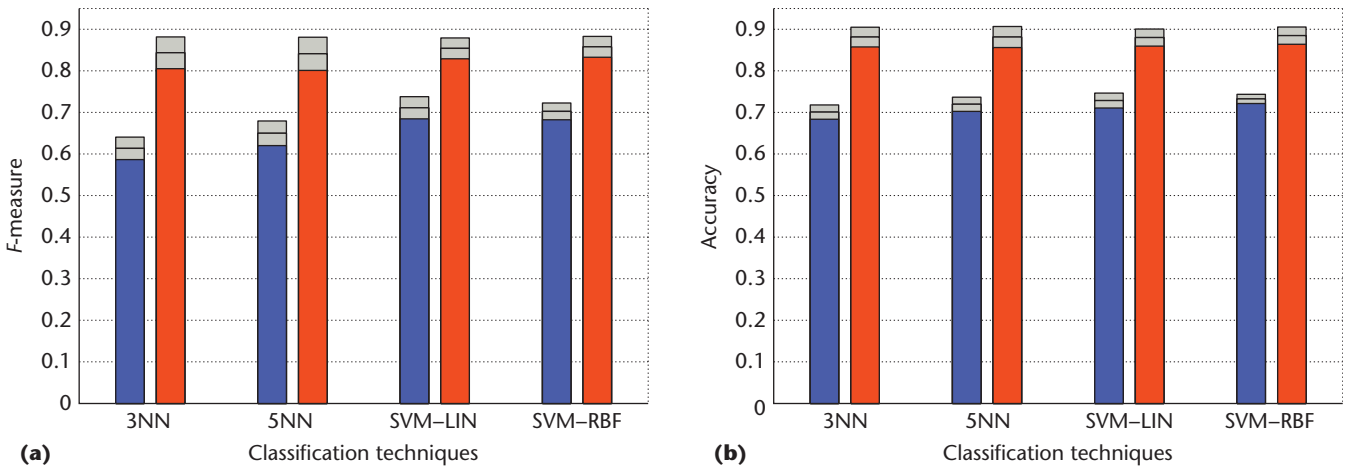

Figure 3. Comparison of classification performance between the feature space of Quack, Leibe, and Van Gool[6] (in blue) and our extension (in red). For each bar, the standard deviation is plotted in gray at the top. Only the (a) F-measure and (b) accuracy diagrams are presented for the 50-50 split. A similar picture holds for the diagrams of precision, recall, and the 66-33 split.

the percentage of images that were annotated as events. In only one cluster did one of the two annotators find that the percentage of landmark images (57.1 percent) was approximately equal to the number of event images (42.8 percent). The same annotator found another cluster consisting of 26.3 percent landmark images and 73.7 percent event images. For the rest of the clusters, both annotators annotated the large majority of images (>80 percent) to belong to only one of the two classes. Thus, it's evident that the vast majority of clusters are pure. That is, they largely contain images of the same class, and can thus be considered suitable for the task of landmark and event classification.

Subsequently, we annotated all 2,056 image clusters derived from clustering the hybrid similarity graph. Each image cluster could be classified as landmark or event, but it was also possible to assign no class to the cluster if the cluster didn't contain images related to some specific entity (a landmark or an event). Out of the 2,056 clusters, 969 were landmarks, 636 were events, and 451 were left unassigned. Subsequently, we trained a set of four variants of standard classification algorithms using landmarks and events as the classes of interest. We left out the unassigned clusters. We used the following four classifier variants: *k*-nearest neighbor with $k = 3$ (3NN) and $k = 5$ (5NN); and SVM with a linear kernel (SVM-LIN) and a radial basis function kernel (SVM-RBF).

We used ten random 50-50 and 66-33 splits of the ground truth in the training and test set and computed mean and standard deviation values for the precision, recall, *F*-measure, and accuracy across the splits. We repeated the experiments for the two-dimensional feature space of Quack, Leibe, and Van Gool, and our extended cluster feature space (incorporating tag information). Figure 3 depicts the results. Careful inspection of the results reveals that

Figure 4. Positions of the top five landmarks as identified by the proposed cluster merging and labeling scheme and five randomly selected events. For each landmark and event, three randomly selected images are shown. The landmark titles were automatically extracted, while the event types were contained as tags associated with images of each respective cluster.

remarkable performance gains can be observed for all classifiers and for both training and test splits thanks to the use of the proposed extended feature space. For instance, there is an increase of almost 23 percent in the *F*-measure of the approach using *k*-nearest neighbor with $k = 3$ for both splits and a corresponding increase in the order of 16 percent for the SVM-RBF classifier reaching an *F*-measure of 87 percent. Therefore, we can conclude that the proposed tag-based features are of particular importance for the success of landmark and event classification.

Following the cluster-processing approach discussed previously, we formed the spatial-proximity graph containing the image clusters corresponding to landmarks. The graph comprises 590 nodes and 10,849 edges. By clustering this graph, we obtained 38 metaclusters. Examination of these metaclusters revealed that 34 of them corresponded to well-known landmarks or points of interest in Barcelona. Five out of the 34 well-recognized metaclusters contained image clusters that didn't correspond to the metacluster landmark (they were placed in the same metacluster due to their spatial proximity with the geographical cluster center). For 29 landmark clusters, for which we could find their actual location on the Web, the automatically generated cluster center fell on average within 80 meters of the actual landmark position. Such accuracy is satisfactory given the fact that some of the landmarks, such as parks, palaces, and so on, span many hundreds of meters. Figure 4 illustrates the location of the top five landmark

Supplementary Video

The video at http://www.computer.org/multimedia/papadopoulos/webextra illustrates a new approach for mining landmarks and events in large, tagged photo collections. Starting from the need for such an approach in today's increasingly media-abundant landscape, the video proceeds through a step-by-step explanation of the proposed approach. It describes the photo similarity graph creation process, the graph-based clustering algorithm, and the photo-cluster classification. Further, the video presents the obtained evaluation results and showcases ClustTour, an online application facilitating the discovery of interesting spots and activities in a travel destination.

Related Work

Landmark and event detection have been usually dealt with as separate problems; for instance, several works[1-3] deal with the problem of landmark recognition, while other works[4-6] address the problem of event detection in social media. Moreover, there have been studies that consider the identification of event and place semantics as parts of the same problem.[7,8]

Landmark detection

The majority of landmark-detection approaches to date exploit the abundance of metadata-rich photos in social-content sources. More specifically, they make use of the visual redundancy in the photo content (that is, many photos that depict the same scene), as well as of the tags and the geolocation information of photos in order to select subsets of images that potentially correspond to landmarks. For instance, one approach[2] first clusters a collection's photos by use of gist descriptors. It then refines the clustering by means of sophisticated geometric verification, and finally uses tag-based filtering to further enhance the quality of the landmark model. However, it neither addresses the problem of identifying image sets that correspond to landmarks (they experiment with landmark-focused data sets) nor does it contain any discussion on the pertinent problem of event detection.

An alternative approach[3] takes a given set of geotagged photos, identifies landmark and geographic tags, then performs a visual clustering of the images tagged with these tags, and finally ranks the derived clusters and the images within them in order to select representative snapshots of the identified landmarks. A similar approach,[1] which apart from a social-content source, uses external online sources (travel guides and image search engines) to enrich the landmark name thesaurus and image collection to be analyzed. Both of these approaches are different from ours, because they perform image clustering only by use of visual similarities and use tag information only for the identification of landmark names. Furthermore, they don't address the problem of event detection.

Event detection

The simplest case for event detection deals with event detection on single images. Joshi and Luo[6] attempt to classify individual images from Flickr into a predefined set of event and activity classes by making use of both visual features and statistical associations between geotags and events. However, they don't exploit the temporal information of images nor do they consider multiple images that correspond to the same event as a single entity. Thus, their results are highly sensitive to noise, which is more conspicuous when media documents are analyzed in isolation.

Other work[4] attempts to identify events through the temporal and spatial features of tag usage. First, event tags are identified by means of a discrete wavelet transform on the temporal and locational distributions. Subsequently, a distinction between periodic and aperiodic event tags is made, and finally image groups that correspond to events are identified on the basis of the image tags. The work presented in the main article is different from certain approaches[4,8] because we directly pursue event identification

metaclusters detected by this step along with three images and a tag automatically selected for each one of them.

Manual examination of the 636 identified event clusters revealed that a large number (43.1 percent) of identified events were related to music (such as concerts and DJ sets). Furthermore, a substantial number of events (9.3 percent) were related to personal events (such as going out with friends). Other important categories of events with respect to their presence in the data set were related to conferences (6.5 percent), traditional and local events (4.6 percent), car and motorbike races (3.3 percent), family occasions (2.9 percent), sailing trips and races (2.8 percent), football matches (2.6 percent), festivals (2.4 percent), expositions (2.3 percent), dancing acts (1.5 percent), and theatrical plays (1.5 percent). Five examples of such events are presented in Figure 4.

Conclusions

Landmark and event detection is a valuable tool for organizing large collections of user contributed images. We have exploited the results of this work in an online travel application for place exploration, named ClustTour.[10] In the future, we plan to investigate the impact of different image similarity graph construction strategies on result precision. For instance, personal image upload and tagging styles will be taken into account when computing pairwise image similarities. Furthermore, we are planning to extend our framework to the analysis of user contributed videos. **MM**

Acknowledgments

This work was supported by the WeKnowIt and Glocal projects, and partially funded by the

on the image domain and not by use of tags. In another study,[5] Becker et al. present a series of experiments where they test different image-similarity learning strategies to cluster event images into sets corresponding to the real-world events they describe. However, they don't consider the problem of distinguishing between event and landmark images, nor do they make use of the visual similarities between images.

Landmark and event detection

There are only few works that consider the problems of landmark and event detection in tandem. In one approach,[8] the authors present a set of methods for associating place and event semantics to tags by means of statistical significance tests that identify temporal and spatial segments with tag usage distributions that are different than expected. Thus, their approach is largely different from ours because they don't directly address the problem of landmark and event detection, focusing instead on tag classification for the classes ''place'' and ''event.''

Quack, Leibe, and Van Gool[7] employ a clustering scheme to group images into coherent clusters by use of both visual and textual (title, description, and tag) features. Then, hierarchical agglomerative clustering is used and a standard classifier is trained to distinguish between objects and events. Finally, the detected objects are mapped when possible to appropriate Wikipedia articles. Our work employs a similar rationale for distinguishing between landmarks and events, but we employ community detection for clustering images and devise a more sophisticated cluster feature space, which leads to better results.

References

1. Y.-T. Zheng et al., ''Tour the World: Building a Web-Scale Landmark Recognition Engine,'' *Proc. Int'l Conf. Computer Vision and Pattern Recognition,* 2009.
2. X. Li, ''Modeling and Recognition of Landmark Image Collections Using Iconic Scene Graphs,'' *Proc. 10th European Conf. Computer Vision,* LNCS 5302. Springer-Verlag, 2008, pp. 427-440.
3. L.S. Kennedy and M. Naaman, ''Generating Diverse and Representative Image Search Results for Landmarks,'' *Proc. 17th Int'l Conf. World Wide Web,* ACM Press, 2008, pp. 297-306.
4. L. Chen and A. Roy, ''Event Detection from Flickr Data through Wavelet-Based Spatial Analysis,'' *Proc. 18th ACM Conf. Information and Knowledge Management,* ACM Press, 2009, pp. 523-532.
5. H. Becker, M. Naaman, and L. Gravano, ''Learning Similarity Metrics for Event Identification in Social Media,'' *Proc. 3rd ACM Int'l Conf. Web Search and Data Mining,* ACM Press, 2010, pp. 291-300.
6. D. Joshi and J. Luo, ''Inferring Generic Activities and Events from Image Content and Bags of Geo-Tags,'' *Proc. Int'l Conf. Content-Based Image and Video Retrieval,* ACM Press, 2008, pp. 37-46.
7. T. Quack, B. Leibe, L. Van Gool, ''World-Scale Mining of Objects and Events from Community Photo Collections,'' *Proc. Int'l Conf. Content-Based Image and Video Retrieval,* ACM Press, 2008, pp. 47-56.
8. T. Rattenbury, N. Good, and M. Naaman, ''Towards Automatic Extraction of Event and Place Semantics from Flickr Tags,'' *Proc. 30th Ann. Int'l ACM Sigir Conf. Research and Development in Information Retrieval,* ACM Press, 2007, pp. 103-110.

European Community, under contract numbers FP7-215453 and FP7-248984.

References

1. S. Fortunato, ''Community Detection in Graphs,'' *Physics Reports,* vol. 486, 2010, pp. 75-174.
2. D.G. Lowe, ''Distinctive Image Features from Scale-Invariant Keypoints,'' *Int'l J. Computer Vision,* vol. 60, no. 2, 2004, pp. 91-110 (software available at http://www.cs.ubc.ca/~lowe/keypoints/).
3. K.E.A. Van de Sande, T. Gevers, and C.G.M. Snoek, ''Evaluating Color Descriptors for Object and Scene Recognition,'' to be published in *IEEE Trans. Pattern Analysis and Machine Intelligence,* 2010.
4. R. Zhao and W.I. Grosky, ''Narrowing the Semantic Gap—Improved Text-Based Web Document Retrieval Using Visual Features,'' *IEEE Trans. Multimedia,* vol. 4, no. 2, 2002, pp. 189-200.
5. X. Xu et al., ''SCAN: A Structural Clustering Algorithm for Networks,'' *Proc. 13th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* ACM Press, 2007, pp. 824-833.
6. T. Quack, B. Leibe, and L. Van Gool, ''World-Scale Mining of Objects and Events from Community Photo Collections,'' *Proc. Int'l Conf. Content-Based Image and Video Retrieval,* ACM Press, 2008, pp. 47-56.
7. X. Li, ''Modeling and Recognition of Landmark Image Collections Using Iconic Scene Graphs,'' *Proc. 10th European Conf. Computer Vision,* LNCS 5302. Springer-Verlag, 2008, pp. 427-440.
8. L.S. Kennedy and M. Naaman, ''Generating Diverse and Representative Image Search Results for Landmarks,'' *Proc. 17th Int'l Conf. World Wide Web,* ACM Press, 2008, pp. 297-306.
9. S. Papadopoulos et al., ''Image Clustering through Community Detection on Hybrid Image Similarity Graphs,'' *Proc. Int'l Conf. Image Processing,* IEEE Press, 2010.

10. S. Papadopoulos et al., ''ClustTour: City Exploration by use of Hybrid Photo Clustering,'' *Technical Demos Program ACM Multimedia Int'l Conf.*, ACM Press, 2010.

Symeon Papadopoulos is a research associate at the Informatics and Telematics Institute, Centre for Research and Technology Hellas, and is also a PhD candidate at the Aristotle University of Thessaloniki. His research interests include Web information mining. Papadopoulos has a professional doctorate in engineering from Technical University of Eindhoven. Contact him at papadop@iti.gr.

Christos Zigkolis is a research assistant at the Informatics and Telematics Institute, Centre for Research and Technology Hellas, and is also a PhD candidate at the Aristotle University of Thessaloniki. His research interests included Web mining and recommendation. Zigkolis has an MS from the department of informatics at Aristotle University of Thessaloniki. Contact him at chzigkol@iti.gr.

Yiannis Kompatsiaris is a senior researcher at the Informatics and Telematics Institute, Centre for Research and Technology Hellas. His research interests include semantic multimedia analysis, social media mining, and the Semantic Web. Kompatsiaris has a PhD in 3D model-based image sequence coding from Aristotle University of Thessaloniki. Contact him at ikom@iti.gr.

Athena Vakali is an associate professor in the informatics department of the Aristotle University of Thessaloniki. Her research interests include Web usage mining, content-delivery networks on the Web, Web data clustering, and Web data caching and outsourcing. Contact her at avakali@csd.auth.gr.

Special Student Offer!

Join IEEE and IEEE Computer Society for only US $40—and receive FREE access to the Computer Society Digital Library (CSDL)

Your benefits include

- Access to FREE development software from Microsoft
- Access to 600 technical books from Safari® Books Online
- Access to 3,500 online Element K® courses available in 10 languages and 1,000 Virtual Labs
- Access to 25+ scholarships

Your CSDL access gives you

- All 27 Computer Society peer-reviewed periodicals with full archives, covering the spectrum of computing and information technology
- 3,800+ conference publications from around the globe
- 380,000+ top quality articles and papers for serious research or quick answers

Join IEEE and IEEE Computer Society today for just US $40 and enjoy benefits to 31 December 2011

Current IEEE students—add Computer Society membership for US $8 (US, Canada) or US $13 (Rest of World)

www.computer.org

IEEE

IEEE computer society

AD-2011STU