# Predicting Actions from Static Scenes

Tuan-Hung Vu, Catherine Olsson, Ivan Laptev, Aude Oliva, Josef Sivic

# Predicting Actions from Static Scenes

Tuan-Hung Vu[1], Catherine Olsson[2], Ivan Laptev[1],
Aude Oliva[2] and Josef Sivic[1]

[1]WILLOW, ENS/INRIA/CNRS UMR 8548, Paris, France,
[2]CSAIL, MIT, Cambridge, Massachusetts, USA

**Abstract.** Human actions naturally co-occur with scenes. In this work we aim to discover action-scene correlation for a large number of scene categories and to use such correlation for action prediction. Towards this goal, we collect a new SUN Action dataset with manual annotations of typical human actions for 397 scenes. We next discover action-scene associations and demonstrate that scene categories can be well identified from their associated actions. Using discovered associations, we address a new task of predicting human actions for images of static scenes. We evaluate prediction of 23 and 38 action classes for images of indoor and outdoor scenes respectively and show promising results. We also propose a new application of geo-localized action prediction and demonstrate ability of our method to automatically answer queries such as "Where is a good place for a picnic?" or "Can I cycle along this path?".

**Keywords:** Action prediction, scene recognition, functional properties

## 1 Introduction

Our environments, such as living rooms, cafes and offices, vary in objects and geometry, but also in *actions* that we usually do in these places (e.g., we typically *work* in offices and *cook* or *eat* in kitchens). As illustrated in Figure 1, scene types are, indeed, often correlated with specific sets of typical actions. The goal of this work is to explore such correlation and to develop algorithms able to answer questions such as "What are typical actions for a given scene?", "Where is a good place to have a picnic?" or "Can I cycle along this path?". Automatic answers to such questions could be useful for several purposes. First, action prediction could provide scene-specific priors when recognizing human actions. For example, relaxing is common on beaches but not on streets; cooking is common in kitchens but not in offices. Second, deviations from an expected set of actions could be used to identify abnormal activities. Third, as we show in this paper, automatic action prediction for geo-localized images could support the search of places suited for particular purposes.

Computer vision has a rich body of work on recognizing human actions [1–5] and scenes [6–9]. Most of this work addresses the problems of action and scene recognition separately. Recently, several methods have shown advantages of recognizing actions or tracking people in the context of their environments [5,

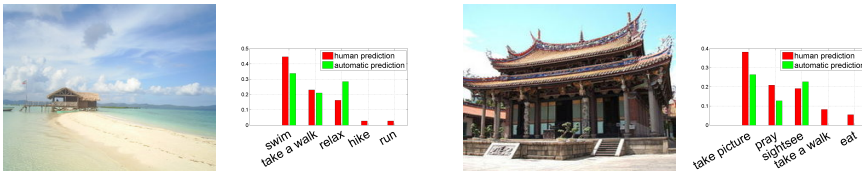



Fig. 1: Images of scene classes `sandbar` and `temple_east_asia` from the SUN dataset [14] together with probabilities for the five most likely actions, predicted manually by people (red) and by our method (green).

10]. Similarly, the interplay between human poses and objects has been studied in [11–13]. While previous work has looked at functional properties for a few selected classes of scenes and objects, here we aim to exploit correlation between scenes and actions at a *large scale* of hundreds of scene categories. Using the discovered correlations, we demonstrate prediction of human actions for test images of outdoor scenes such as, for example, found on Google maps.

To reach our goal, we construct a new SUN Action dataset and collect manual annotations of human actions for 7940 images of 397 scene categories from the SUN dataset [14]. Analysis of this data reveals strong action-scene correlation for the majority of scene categories. Notably, we show that an image's scene category can be determined from corresponding textual descriptions of characteristic actions for that image.

Using the discovered relations between scenes and actions, we next address the task of automatic action prediction for images of static scenes. We consider 38 outdoor and 23 indoor action classes and associate typical action labels with 397 scene categories. Using such scene-based action annotation we learn visual classifiers for each action category and predict actions for images of static scenes as illustrated in Figure 1.

We finally demonstrate an application of our method to the new task of geo-localized action prediction. Our motivation comes from the large amount of publically-available geo-tagged images (e.g., on Flickr and Panoramio) which is expected to grow even faster with the introduction of new wearable devices such as Google Glass. Application of automatic action prediction to such images will enable the search for places based on their *function*, including specific actions such as swimming, having picnic, hiking and many others. In our experiments we use geo-localized images of outdoor scenes collected from panoramio.com and demonstrate examples of successful action prediction on the map of France.

In summary, we make the following three contributions. First, we present a new dataset with manual annotations of typical actions for 397 scene classes (see Section 3) and use it to analyze action-scene correlations (see Section 4). Second, based on the discovered correlations, we demonstrate successful action prediction for images of static scenes (see Section 5). Finally, we propose a new task of geo-localized action prediction. We apply our method to geo-tagged images on the web and show encouraging results of searching maps for locations suitable for particular activities (see Section 6).

## 2 Related work

Relatively few papers explore relations between scenes and actions. Li *et al.* [15] propose a graphical model combining evidence from object and scene categories for action recognition in still images. Marszalek *et al.* [5] propose a joint framework for scene and action recognition in video. While most of the work in action recognition targets actions depicted in images or video, here we address a different task and predict actions in scene images with no action observations.

Action prediction has been recently addressed by Kitani *et al.* [10] and Walker *et al.* [16] aiming to model future motion of people and cars using priors derived from the scene. Yuen and Torralba [17] predict motion for images of static scenes by searching and transferring motion cues from video scenes with similar appearance. Our work complements these efforts and investigates action prediction for a large set of scenes and actions.

Recognition of functional properties of objects and scenes is an interesting but less explored area of computer vision. Relations between people and objects as well as between human poses and scene geometry have been investigated in [11–13]. Patterson and Hays [18] annotate scene images with a set of global attributes of various types (i.e: material, surface property, affordance and spatial envelope), and recognize attributes from scene images. Unlike any previous work, we here aim to model functional properties for a wide range of scene classes. Our work is similar in spirit to Arietta *et al.* [19] and Khosla *et al.* [20] who aim to predict non-observed scene properties such as crime rate in the area.

## 3 Dataset

**Dataset annotation.** To analyze correlations over a wide range of scene categories and a rich set of actions, we gather the novel *SUN Action* dataset (short for "Scene UNderstanding - Action") with manual annotations of typical human actions for images of static scenes. We use scene images from the SUN dataset [14]. For each of the 397 well-sampled scene categories we collect free-form annotations of typical actions for the twenty "most typical" images in that category [21], for a total of 7940 images. Annotations were crowdsourced using Amazon Mechanical Turk (AMT)[1]. AMT workers were shown images of scenes and were asked to list between one and three words or short phrases for each scene describing a typical action that one would usually do there. Scene category labels were not provided. All together we collected 137,558 responses: each image received 17.3 responses on average, and each category received an average of 346.4 responses.

Example images and corresponding responses from the SUN Action dataset are shown in Figure 2. We have observed a varying diversity of responses for different scene categories. The top row of Figure 2 shows a few examples of scene classes with low entropy of response histograms (high annotator agreement,

[1] AMT workers gave consent (set by the MIT IRB) for each HIT they chose to perform.

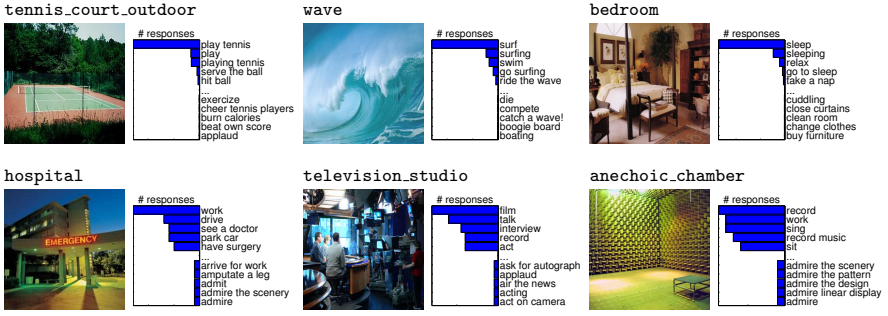



Fig. 2: SUN Action scene categories with corresponding histograms of action responses. Top row: Scene categories with low entropy of response histograms. Bottom row: scene categories with high entropy of response histograms. Low-entropy categories are often places designed for specific purposes (tennis court) or where the environment limits possible actions (wave). By comparison, high-entropy categories are places that afford many actions (television studio) or are unfamiliar (anechoic chamber)


Fig. 3: Histograms of words in action responses for two images of the scene class `crosswalk`. The presence of a cyclist in the image on the right biases responses to contain the action "bike", which is not present in other crosswalk images.

low response diversity). Such scenes often correspond to places that have been designed for specific purposes (tennis court) or where the natural environment limits the set of possible actions (wave). In contrast, scene classes with high entropy of responses (Figure 2, bottom) are places that afford many actions (e.g., a television studio, where many actions need to take place over the course of filming) or unfamiliar places (an anechoic chamber).

The majority of images in the SUN Action dataset contain no people. We found this property to be important for collecting unbiased annotations of typical actions. For a few images containing people we have observed a bias in action annotations towards actions depicted in the image. An example of such a bias is shown in Figure 3 illustrating two crosswalk scenes, one without people and one with a cycling person. In the scene containing the cyclist, the predominant response was "bike", unlike other images in the crosswalk category.

**Processing of action responses.** Action responses were gathered in free-form natural language and require preprocessing for our further analysis. Many of responses contain nearly identical information but differ in grammatical structure, such as "read the book while on the flight" and "read a book". Our first pass of preprocessing converts responses into simplified action annotations by extracting verbs or verb-noun patterns from each response. This strategy reduces the response space while preserving meaning. For example, responses like *"read the*

*book while on the flight"* or *"avoid eye contact with neighbours"* are trimmed to *"read book"* and *"avoid eye contact"* respectively. We use the Stanford NLP toolbox [22] for part-of-speech tagging, stemming, and removal of stop words, and extract either verbs or verb-noun patterns from each response. Responses containing no verbs are removed. The words extracted in this stage of preprocessing are used as input to predict scene categories in Section 4.

For the action prediction task in Section 5 we manually group semantically similar action responses into action classes. To define action classes, we automatically extract 100 most frequent verb patterns, i.e. single verbs, verb+noun, etc., from action responses. Patterns with similar meaning are then manually merged yielding action labels, for example, "walk on grass" and "walk on sand" are merged into "walk". We note that the automatic parsing of natural language into action categories is an open problem beyond our work. In particular, we separate scenes into 197 outdoor and 203 indoor categories and define corresponding 38 outdoor and 23 indoor action classes as listed in Figure 7.

Given the average of 17.3 action responses per image in our database and a potentially large number of typical actions for a scene, our per-image annotation is not exhaustive. To address this problem, we assume that instances of the same scene category share the same functional properties. We found this assumption to be valid in most cases in our database. We therefore assign the same action labels to all instances of a given scene category using the following *label propagation* strategy. A scene category $C$ is labeled by an action $A$ if images of $C$ are labeled with $A$ at least 20 times. Following this procedure, for each action label $A$ we obtain a set of *positive* scene categories. The *negative* scene categories for $A$ are those containing no $A$ labels for any of their images. Results of our preprocessing together with the original action responses are available from [23].

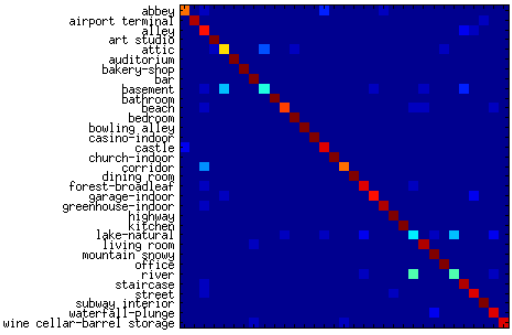
## 4 Analysis of scene-action correlation

Are different scene categories correlated with distinctive sets of actions? Scene categories are often defined by what you would typically do there: for example, in an office one would typically *work*, whereas in a kitchen one would typically *cook*. Indeed, most man-made scenes around us have been created to facilitate certain actions.

This section verifies and quantifies relations between actions and scenes. We demonstrate successful recognition of a large number of scene categories from associated actions descriptions. We further investigate the structure of action-scene correlations with a hierarchical clustering analysis.

### 4.1 Predicting scenes from actions

To verify the hypothesized correlation between scene categories and actions, we conduct two classification experiments using action annotations in the SUN Action dataset. We take inspiration from the field of text classification. In the SUN Action dataset, each image is associated with a collection of natural-language




(a)

| Method | 33-cat | 397-cat |
|---|---|---|
| Chance level | 3.00 | 0.25 |
| Nearest Centroid | 85.80 | 40.31 |
| ML Naive Bayes | **91.97** | **55.86** |

(b)

Fig. 4: Results of action-based scene classification. (a): Confusion matrix for the 33-category subset using Maximum Likelihood Naive Bayes estimation. The strong red line along the diagonal indicates excellent classification performance. A few pairs of categories e.g., (basement,attic) and (river,lake) are confused due to similarity in their characteristic actions. (b): Average accuracy (%) of scene classification for the 33-category subset and for all 397 scene categories.

action descriptions. Classifying images based on a collection of associated responses is reminiscent of classifying documents based on their contents. However, there are two notable differences in our approach. First, the number of responses available per image (17.3 responses on average) is significantly lower compared to the number of words in a typical text document. Secondly, we wish to probe category membership using only a small collection of responses per image, to simulate asking a handful of people to provide a most typical action for the image and then performing classification based on the consensus of that set of responses. Therefore we use classification strategies that compare small queries to entire categories at a time.

**Classification methods.** We classify images using two simple bag-of-words techniques – Nearest Centroid and Naive Bayes. First, we divide the images in each class into 10 folds for cross-validation. Within the training set, the responses for each image are split into individual words. These word counts are combined and normalized across all images within a given class, to generate a word distribution histogram for each scene category. Within the test set, responses for each image are randomly grouped into chunks of 7 responses for that image, to simulate asking a handful of people at a time to provide a most typical action for each image. Responses within each chunk are then split into individual words to form bag-of-words queries.

In nearest centroid classification, the bag-of-words queries are normalized to form histograms, which are compared with category histograms according to histogram intersection distance. The scene category centroid with the smallest distance from the query is selected as the class label.

In Maximum Likelihood Naive Bayes classification, the category histograms are interpreted as empirical likelihood estimates: the likelihood $\Pr(w|c)$ of observing word $w$ in association with an image of class $c$ is assumed to be the

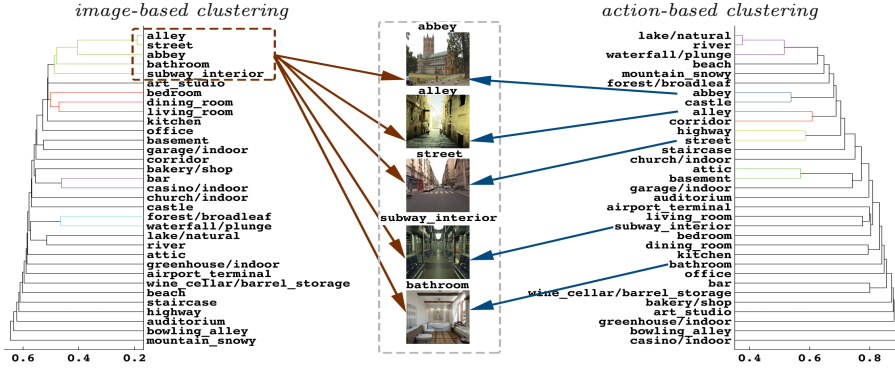



Fig. 5: Results of hierarchical clustering of 33 scene categories based on the similarity of image descriptors (left) and action similarity (right). Image-based similarity groups similar-*looking* scenes despite their large difference in semantics such as "alley" and "bathroom". In contrast, action-based similarity results in more semantically meaningful clusters. For example, "mountain, snowy" is placed in a category of its own according to the visual similarity, whereas it is grouped together with other outdoor places on the basis of action similarity.

number or observations of $w$ within the class $c$ responses in the training set divided by the total number of words in all class $c$ responses combined. The word observation likelihoods are assumed to be conditionally independent (the "naive Bayes assumption"), enabling us to compute the class-conditional likelihood of each bag-of-word query as the product of each constituent word's empirical class likelihood: $\Pr(w_1, w_2, \ldots w_n|c) = \Pr(w_1|c) \times \Pr(w_2|c) \times \ldots \times \Pr(w_n|c)$. The empirical likelihood estimate makes no explicit provision for estimating the likelihood of unobserved word-class pairs. To address this issue, we compute the minimum class-conditional likelihood over all words and classes in the dataset, $\min_{w,c}(\Pr(w|c))$, and use this probability to stand in as the class-conditional likelihood for unobserved words. We assume a uniform prior over scene categories, enabling maximum likelihood estimation: that is to say, bag-of-words queries are classified according to which class provides the largest class-conditional likelihood.

**Results.** Figure 4 illustrates results of scene classification. As visualizing results across 397 individual classes is difficult, we select a 33-category subset of well-recognized and semantically important scene categories. To select the 33-category subset, we have asked four of our collaborators to nominate 20-40 most important scene types. Out of the 80 scene types with most annotated images, 35 received at least two nominations and were slated for inclusion. "Cathedral" was removed for not being different enough from "church", and "abbey" and "coast" was removed for containing only aerial shots, leaving a final slate of 33 scene categories.

The confusion matrix for a Naive Bayes method in Figure 4(a) shows a strong diagonal indicating excellent classification performance. While most classes have

almost perfect classification accuracy, a few classes are confused by the classifier due to the sharing of common actions. For example, scene categories "basement" and "attic" are both often annotated by actions "store" and "clean", while scene categories "river" and "lake" are frequently labeled with "swim" and "fishing".

Quantitative classification results of the two methods for the 33-category subset and all 397 scene classes are shown in Figure 4(b). Notably, both methods perform considerably better than chance while Naive Bayes provides better performance than Nearest Centroid. The fact that such simple classification methods yield very good performance indicates a strong correlation between scene categories and human actions: different scene categories have distinct patterns of associated actions. This confirms our initial hypothesis of a very strong relation between scene categories and their functional properties.

### 4.2 Action-based scene clustering

We seek to further investigate the structure of correlations between scene categories and actions: Which scene categories are more similar in terms of their function? We use hierarchical clustering and group scene descriptors at multiple scales. At the finest scale, only the most similar scene types cluster together, whereas at coarse scales, clusters are larger and encompass more dissimilar scene types. Dendrogram visualizations in Figure 5 show the progression of clustering patterns from fine to coarse: categories are represented as "leaves"; branchings closer to the leaves of the tree connect classes that cluster together under fine-grained clusterings; and branchings closer to the trunk of the tree encompass broader clusterings. The height of each linkage in the dendrogram indicates the distance between the subclusters it connects.

The two dendrograms in Figure 5 illustrate image-based and action-based scene clustering. In the first case, distances between scenes were obtained as Eucledian distances of corresponding image descriptors (see Section 5). Clustering based on human action annotations was obtained using $\chi^2$-distances between scene representations in terms of bag-of-words histograms (see Section 4.1). We observe that image-based similarity in Figure 5(left) captures substantially different information about scene classes as compared to action similarity Figure 5(right). For example, "alley", "bathroom" and "subway interior" are grouped together according to visual similarity due to their similar geometrty and texture, but are separated according to action similarity since alleys, bathrooms and subways have different function. Another example is that visual similarity places "mountain, snowy" is a category of its own because no other class commonly depicts open white peaks, whereas action similarity places mountains together with other outdoor places that are associated with hiking, taking photos, and related actions.

## 5 Visual Action Prediction

People can easily determine appropriate actions to perform in a given place. Are machines able to do the same thing? In Section 4 we have addressed the related

problem of predicting scene categories from a set of associated actions. Here we turn to the problem of predicting typical actions for an image of a scene. We approach the problem of visual action prediction using standard image classification techniques in terms of local features and binary classifiers. To train image classifiers we use action labels derived from action annotations as described in Section 3. We predict actions separately for indoor and outdoor scenes.

We test two different schemes for action prediction. Under the first scheme (**S1**), we train action classifiers directly from images using action labels only. Under the second scheme (**S2**), we first classify images into scene categories as an intermediate step and then assign action scores based on the obtained scores of scene classifiers. We assume that any particular test image belongs to one scene category only, therefore a score for an action $a$ in a given image is defined as a max scene score over scene categories $S$ associated with $a$.

### 5.1 Implementation Details

**Image representation.** Our image classification pipeline follows standard approaches and consists of densely extracted local image features, a learned visual vocabulary and a feature encoding step. For local image features we use HOG2x2 [14], SIFT [24] and CSIFT [25, 26] descriptors. Descriptor dimension is reduced by PCA. For the encoding phase we consider two popular encoding techniques: histogram encoding (BoW) and Fisher Vector encoding (FV). To exploit spatial information, we apply Spatial Pyramid framework [6], using grids of size 1x1, 2x2, 3x1. Each grid cell is represented either by BoW or FV vectors. The resulting vectors are normalized and concatenated to create the final representation. In the rest of this section we use the format *<descriptor>_<encoding technique>* to denote image representation techniques, e.g. CSIFT_FV as Fisher Vector encoding for CSIFT descriptors.

**Classification.** For the classification, we train SVM classifiers using LIBSVM toolbox [27]. Linear kernels are used for image representation by Fisher Vector. For the histogram representation (SIFT_BoW /CSIFT_BoW), we use $\chi^2$ kernel [28]. With HOG_BoW, we exploit Histogram Intersection kernel. Training by SVM, we can boost up the performance by simply using a linear combination of kernels. In our experiments, we aggregate kernels with equal weights.

### 5.2 Experimental results

For training and testing the classifiers, we randomly divide the dataset into two equal parts. Our training and testing splits are balanced in number of images per scene category. Our results for action prediction are summarized in Figures 6-8 and Table 1.

We use mean Average Precision (mAP) as the performance measure. To get mAP, we first compute the area under precision-recall curve, or Average Precision (AP), for each class. Then mAP is determined as the mean of average


Fig. 6: Automatic visual action prediction for test images in SUN Action dataset.

precisions across all classes. We obtain best prediction mAP of 60.99% for outdoor actions and 52.09% for indoor actions using combination of HOG_BOW, HOG_FV and CSIFT_FV kernels. Our result is significantly higher than the mAP at chance level, i.e. 6.32% for outdoor action classes and 4.24% for indoor action classes. Figures 6-8 are produced with our best kernel combination.

In Figure 7, we show classification results with 38 outdoor and 23 indoor action classes sorted by AP. For better visualization of prediction results for some action classes, we show example images in Figure 8. The last two columns depict some hard positive and hard negative samples for each class. For outdoor scenes, action classes such as "hike", "pray", having rather typical color/structure, are easier to classify than other "can-do-almost-every-where" action classes like "learn". While people can often differentiate universities from other buildings based e.g., on the text and other cues, the task is still difficult for current vision systems, especially, for those exploiting global image representations. We notice that indoor actions are more structure-dependent than outdoor actions. In our experiment APs of indoor action classes are generally lower than APs of outdoor action classes. We also observe different levels of difficulty among indoor actions, e.g., detecting bowling lanes is easier than detecting sink-like structures. We found building action classifiers challenging, because positive samples are possibly images from very different scene categories, thus covering much larger range of visual texture and structure.

We also aggregate predicted action scores for test images and try to estimate the score contribution. Figure 6 shows some test images along with manual action annotations and automatic action predictions. For this visualization, we map SVM scores of test images to probabilities using Platt's sigmoid [29], with parameters estimated during the training phase. Even though the results are not perfect, we still observe a good match between distributions of annotated and predicted actions. Our predictors successfully give reasonable responses like "swim", "take a walk" and "relax" to a beach image, or "take picture", "pray" and "sightsee" to a temple image. Other qualitative results of action prediction by our method are available at [23].

For more quantitative analysis, we now consider Table 1. The table shows action prediction mAPs of two proposed training schemes combined with different image representation techniques. By comparing results of the two schemes, as shown in two columns **S1** and **S2**, we can conclude that learning action classifiers directly achieve better prediction performance than aggregating mul-

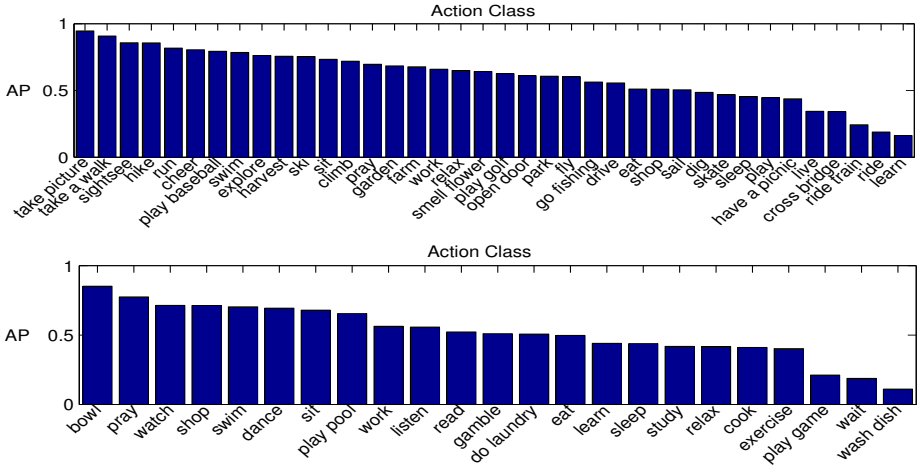



Fig. 7: Results of action prediction for all 38 outdoor actions (top) and 23 indoor actions (bottom) sorted in the decreasing order of Average Precision (AP).


Fig. 8: Selected SUN Action classification results - both outdoor (cyan) and indoor (orange) - with our best kernel combination.

tiple scene classifiers. This improvement can be attributed to sharing similar functional properties across different scene classes. In terms of image representation techniques, Fisher Vector encoding yields better performance compared to Histogram encoding. These results are consistent with recent works on scene

| Method | S1 | S2 | S1-S2 | S1 | S2 | S1-S2 |
|---|---|---|---|---|---|---|
| SIFT_BoW | 40.92 | 40.68 | 0.24 | 31.75 | 28.71 | 3.04 |
| SIFT_FV | 41.15 | 34.51 | 6.64 | 31.04 | 27.13 | 3.91 |
| CSIFT_BoW | 47.78 | 44.43 | 3.35 | 32.53 | 27.90 | 4.63 |
| CSIFT_FV | 49.52 | 41.65 | 7.87 | 36.29 | 29.70 | 6.59 |
| HOG_BoW | 47.03 | 45.93 | 1.10 | 37.91 | 35.50 | 2.41 |
| HOG_FV | 52.66 | 47.75 | 4.91 | 42.78 | 43.89 | -1.11 |
| HOG_BoW+ SIFT_FV+ CSIFT_FV | 56.60 | 50.06 | 6.54 | 46.11 | 40.04 | 6.07 |
| HOG_BoW+ HOG_FV+ CSIFT_FV | **60.99** | 54.25 | 6.74 | **52.09** | 45.98 | 6.11 |
| SIFT_FV+ HOG_FV+ CSIFT_FV | 56.48 | 49.61 | 6.87 | 45.76 | 41.41 | 4.35 |

Table 1: Scene-based outdoor (cyan) and indor (orange) action prediction results with different approaches. Note that mAP at chance level is 6.32% (outdoor) an d 4.24% (indoor). **S1** and **S2** columns respectively show classification mAP (%) of the two aforementioned training schemes. Column (**S1-S2**) shows the different mAP between two schemes. We observe consistently better performance of scheme **S1**: directly training binary action classifiers over scheme **S2**: aggregating scene classifiers.

classification [30]. Significant performance difference between SIFT and CSIFT proves that color information is useful for the task. Also, linear combination of multiple kernels does improve the performance. Among our three tested kernel combinations, using HOG_BoW, HOG_FV and CSIFT_FV yields the best result. In conclusion, we have shown high accuracy for a new task of action prediction evaluated on a large number of action and scene classes.

## 6 Image-based Geo-Mapping of Actions

One possible application of scene-based action prediction is to search for places in which to do a specific action. For example, a user may ask "Where can I camp in the Mont Blanc valley?" or "Where can I sunbathe in Tuscany?". Such queries are currently not supported by map services such as Google Maps or Bing Maps. To address this problem, we introduce Image-based Geo-Mapping of Action (IGMA), an application for geo-localizing actions on a map and answering map queries of the type "Where can I do X?". Results are derived from geo-localized scene images publicly available on the Internet. IGMA is the first attempt to automatically answer geo-localized action queries. Our strategy of predicting actions at a broad spatial scale using geo-localized images enables us to go beyond manual location-action labels: for example, one can "have a picnic" not only at a designated picnic area, but also in a grassy countryside field.

**Collecting the Panoramio dataset.** We use Panoramio image sharing service [31] to collect a dataset with geo-localized images. Like the SUN Action dataset, the images in Panoramio contain few to no people. The Panoramio service provides a REST API for selecting images: given a range of longitude and latitude values, Panoramio returns a JSON file of image properties including im-

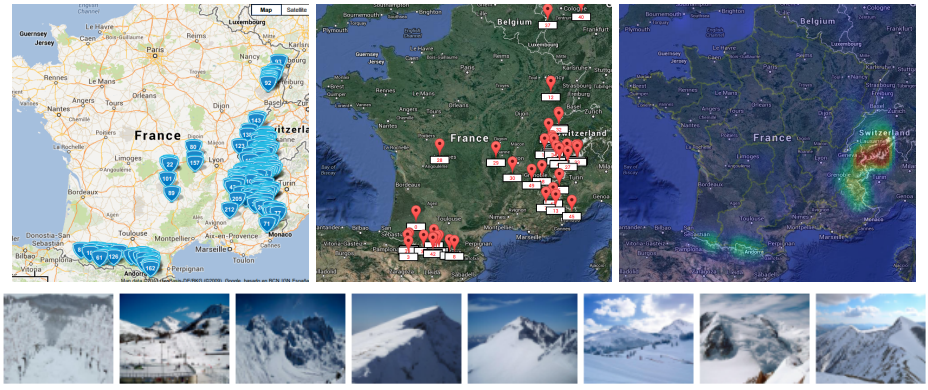



Fig. 9: "Where can I ski in France?" - (Top left) Official skiing stations in France [32]. (Top middle) Suggested places for skiing by IGMA. (Top right) Dense map of action "ski" generated by IGMA. (Bottom) Panoramio images of suggested places for skiing.

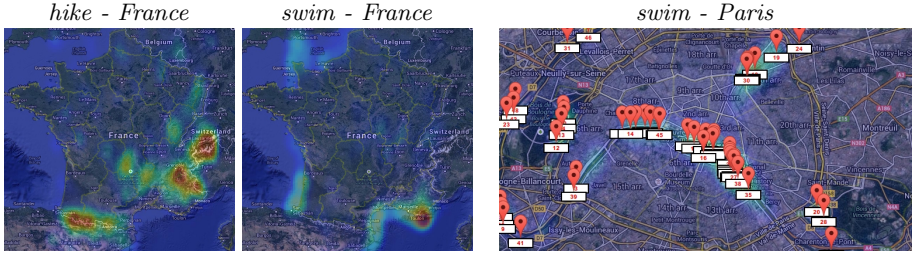



Fig. 10: Geo-localized prediction of actions. (left): Predictions for actions "hike" and "swim" on the map of France. (right): Predictions for the action "swim" in Paris.

age URL and geographical position. For our experiment, we collected Panoramio images of France, with longitude from $-5°$ to $8°$ and latitude from $41°$ to $51°$. In total, our dataset contains over $38,000$ distinct geo-tagged images.

**Dense map of actions.** Our goal is to construct dense maps visualizing places where people would likely perform certain actions. We construct these maps by applying the scene-based action classifiers computed in Section 5.2 to the collected Panoramio dataset using the following procedure.

For a given action, we compute the top-scored Panoramio images. We generate a dense map from this list of scores and geo-locations by modeling the map using a Gaussian Mixture Model (GMM) with mixture components centered at the image locations and their weights set to corresponding action scores. The standard deviation $\sigma$ for each component is set to a fixed value.

This initial dense map estimate is adjusted to compensate for non-uniform sampling of Panoramio images. Different population densities of the examined regions may introduce bias to the action density estimation. Therefore, we estimate the sampling density of Panoramio images. To estimate sampling density, we use the same GMM model above with the same $\sigma$ for each Gaussian compo-

nent as before, but with a uniform weight across all mixture model components rather than an action-score-based weight. The initial action map estimated from the highest scored images is normalized by the estimated sampling density of Panoramio images to correct the sampling density bias. We then get the final estimation of action density.

Figure 9 illustrates IGMA's suggestions for the question "Where can I ski in France?". We compare the estimated dense map produced by IGMA for the action "ski" with the the map of official skiing stations in France, acquired from [32]. Visually, our predictions have a high degree of correspondence with locations containing official skiing areas. Similarly, in Figure 10(left) we illustrate predictions for "hike" and "swim" in France. These results visually correspond to the sea-coast and mountain areas of France, confirming good geo-localization of actions.

Figure 10(right) illustrates an interesting result of predicting the "swim" action in Paris. This result suggests an area for further investigation: the recommended locations for swimming in Paris fall mainly along the river Seine, where swimming is very uncommon. While it is true that scene categories often have strong correlation with associated actions, not all scenes within a scene category share the same action affordances in practice. One possible approach to this issue might be to subdivide scene categories according to more fine-grained functional affordances, e.g. separately identifying rivers where people can and cannot swim.

## 7 Conclusion

In this work we have addressed a new problem of action prediction for a wide range of scene images. We have collected a new SUN Action dataset with manual annotations of typical actions for scene images, and discovered strong action-scene correlation for the majority of scene classes. Based on this correlation, we have learned to predict typical actions for a large set of scenes. Using standard state-of-the-art image classification techniques we have shown high accuracy of action prediction, which is an encouraging result for a new problem. To demonstrate potential advantages of our work, we have shown promising results on a new application Geo-Mapping of Actions (IGMA) enabling automatic answers to queries such as "Where can I do X?".

## Acknowledgements

This work is partly funded by ERC Activia, US National Science Foundation grant 1016862, Google Research Awards, MSR-INRIA laboratory and EIT-ICT labs.

## References


1. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008)
2. Niebles, J., Chen, C.W., Fei-Fei, L.: Modeling temporal structure of decomposable motion segments for activity classification. In: ECCV. (2010) 392–405
3. Sadanand, S., Corso, J.: Action bank: A high-level representation of activity in video. In: CVPR. (2012)
4. Wang, H., Kläser, A., Schmid, C., Cheng-Lin, L.: Action Recognition by Dense Trajectories. In: CVPR. (2011)
5. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR. (2009)
6. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006)
7. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV **42** (2001) 145–175
8. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: CVPR. (2009)
9. Vogel, J., Schiele, B.: Natural scene retrieval based on a semantic modeling step. In: Image and video Retrieval. (2004) 207–215
10. Kitani, K.M., Ziebart, B.D., Bagnell, J.A., Hebert, M.: Activity forecasting. In: ECCV. (2012)
11. Grabner, H., Gall, J., Van Gool, L.: What makes a chair a chair? In: CVPR. (2011)
12. Gupta, A., Satkin, S., Efros, A., Hebert, M.: From 3d scene geometry to human workspace. In: CVPR. (2011)
13. Delaitre, V., Fouhey, D., Laptev, I., Sivic, J., Gupta, A., Efros, A.: Scene semantics from long-term observation of people. In: ECCV. (2012)
14. Jianxiong, X., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR. (2010) 3485–3492
15. Li, L.J., Fei-Fei, L.: What, where and who? classifying events by scene and object recognition. In: ICCV. (2007)
16. Walker, J., Gupta, A., Hebert, M.: Patch to the future: Unsupervised visual prediction. In: CVPR. (2014)
17. Yuen, J., Torralba, A.: A data-driven approach for event prediction. In: ECCV. (2010)
18. Patterson, G., Hays, J.: Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: CVPR. (2012)
19. Arietta, S., Agrawala, M., Ramamoorthi, R.: On relating visual elements to city statistics. Technical Report UCB/EECS-2013-157, EECS Department, University of California, Berkeley (Sep 2013)
20. Khosla, A., An, B., Lim, J., Torralba, A.: Looking beyond the visible scene. In: CVPR. (2014)
21. Ehinger, K.A., Xiao, J., Torralba, A., Oliva, A.: Estimating scene typicality from human ratings and image features. (2011)
22. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. (2003) 173–180
23. http://www.di.ens.fr/willow/research/actionsfromscenes
24. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2) (2004) 91–110

25. Abdel-Hakim, A.E., Farag, A.A.: Csift: A sift descriptor with color invariant characteristics. In: CVPR. (2006)
26. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Empowering visual categorization with the gpu. IEEE Transactions on Multimedia **13**(1) (2011) 60–70
27. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3) (2011) 27:1–27:27
28. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. IJCV **73**(2) (2007) 213–238
29. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in large margin classifiers, MIT Press (1999) 61–74
30. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: CVPR. (2007)
31. Google: Panoramio service. http://www.panoramio.com (2007)
32. : Map of ski stations in france. www.skiinfo.fr/france/carte.html (2013)