Abstract—This paper explores the problem of large-scale automatic video geolocation. A methodology is developed to infer the location at which videos from Anonymized.com were recorded using video content and various additional signals. Specifically, multiple binary Adaboost classifiers are trained to identify particular places based on learning decision stumps on sets of hundreds of thousands of sparse features. A one-vs-all classification strategy is then used to classify the location at which videos were recorded. Results demonstrate that high accuracy video geolocation is indeed possible for many videos and locations and interesting relationships exist between between videos and the places where they are recorded.

I. INTRODUCTION

For many videos, the geographic location at which the video was recorded is a defining characteristic. Indeed, for some, such as tourist videos, the video’s geolocation is paramount. A video is often best described through naming the location that it depicts. Thus, a description or annotation of a video is often incomplete without a reference to a geographic location. Such geographic annotations can offer great utility and there are innumerable reasons why people would be interested in identifying videos from a specific location.

Unfortunately, most videos are not explicitly annotated with a geographic location, either as GPS coordinates or textual metadata. However, one can often infer a video’s geolocation through analyzing the video’s content and metadata. Consider the frames taken from the Anonymized.com videos displayed in Figure 1. From these frames, what can you say about the location at which the video was taken? (a) can be pinpointed to Paris, France by anyone who is familiar with the famous landmark depicted in the image, the Eiffel Tower. One can deduce that (b) depicts a ski slope, thus eliminating well over 99.9% of the earth’s surface, but establishing which ski slope it is or even which continent it is on is difficult without extra information. (c), which depicts divers diving underwater with a whale shark, is even more challenging, as water covers the majority of the earth’s surface. How can one establish the locations depicted in (b) and (c)? Fortunately, text metadata affords some reprieve. The title for (b) is “[Name] boarding in Whistler” and the description for (c) contains the words “in den Malediven” or “in the Maldives” in German. At first, glance (d) seems to be recorded in Paris, France. However, upon closer inspection one may notice that it is in fact a small scale replica of Paris in Las Vegas, Nevada, USA. These examples demonstrate that video geolocation is a challenging task where multiple sources of information must be considered. Furthermore, the utility of various signals differs significantly for various locations. For example, the observation of snow and the words “ski” and “boarding” will be useful to geolocate (b), but will offer very little help with (c).

This paper introduces a methodology for automating the video geolocation process for an immense database of videos from the video sharing site Anonymized.com. In particular, a data set of 20 million labeled videos is used to train classifiers to determine which combinations of features identify particular locations in videos. These classifiers are designed to be applied to a corpus of videos that is significantly larger than the training set. Thus, the emphasis is placed on quick and efficient geolocation at high precision with a low false positive rate.
A. Previous Work

To our knowledge, there exists thus far no literature on the problem of geolocating videos. This is likely due to a lack of labeled data as no standard has been set for annotating videos with geographic locations (such as the EXIF standard for images). However, the literature on image geolocation sets a precedent for videos and will be reviewed here.

The availability of large online photo collections, paired with a universal standard for storing corresponding GPS coordinates has facilitated research in image geolocation. Numerous image based location recognition algorithms employ viewpoint and scale invariant local feature detectors to match photos of visually distinct locations (e.g. [1]–[4]). As the resulting massive feature vocabularies present a challenge for storage and retrieval, [3], [4] develop data structures and algorithms for quick and efficient retrieval. This challenge is, however, significantly compounded for videos as each second of video consists of numerous images requiring a much larger dictionary.

Hays and Efros [5] explored the problem of image geolocation using only the image content and showed that there is a strong correlation between the visual similarity of images and their location. Through finding the most visually similar images or image clusters, using various visual features, they demonstrated that one can significantly narrow down the potential geographic locations at which an image could have been taken. However, the limitations of using only visual features are clear through examples such as “Spaghetti Westerns” – films that depict the American Southwest but are filmed in visually similar regions of Spain. Many locations can be geographically distant but look very similar, such as urban areas, deserts and beaches. In these cases, visual features can be used to reduce the search space of potential locations, but additional information will be needed to specify the exact location.

In addition to using visual features, Kalogerakis et al. [6] predict the geolocation of images by incorporating into their model a prior on human travel patterns. They demonstrate that one can geolocate temporal sequences of photos with higher accuracy than single images by incorporating a spatio-temporal model of how people travel. Chen and Grauman [7] adopt a similar strategy, using timestamped sequences of photos. Through the assumption of spatio-temporal smoothness they are able to identify the locations of “bursts” of photos that correspond to common paths that people traverse.

Van Laere et al. [8] and Serdyukov et al. [9] show that textual annotations associated with photos can be a powerful indicator of geographic location. Zheng et al. [10] developed a web-scale landmark detection engine that detects landmarks in photos by matching visual features to tightly clustered geolocated images. Thus, photos can be geolocated with very high accuracy if they contain visually distinct objects that have been frequently photographed.

Previous studies on the geolocation of images suggest that a number of signals can and should be used to determine the geolocation of a video. Some videos will reveal their location through a clear view of a unique landmark and for some a user may simply state the location in the title of the video. Other videos will require a careful weighting of various weak signals such as a mention of the beach in a particular language, the observation of palm trees and perhaps a particular shade of light blue water.

II. Data Set

When uploading a video to Anonymized.com, uploaders may optionally specify the location at which the video was recorded. The location is either inputted as a text string indicating, for example, a city and country, or as a latitude-longitude pair chosen by clicking on a map. The resulting accuracy of the labeled data depends on how accurately the uploader-supplied location can be resolved. A video that is pinpointed on a map yields latitude-longitude coordinates, the accuracy of which depends on the zoom level and discretion of the user. Inputted text strings can be resolved to various granularities; “Paris, France” can be mapped to a city, while “TX, USA” can be resolved to a large geographic area consisting of the U.S. state of Texas. Some user-indicated locations are ambiguous. The location “Cambridge,” for example, can be a location in England, the U.S., or Canada. Such ambiguous labels are discarded when possible. Unfortunately, only a small fraction of videos are annotated with this information. A subset of approximately 20 million videos, for which the Anonymized.com uploaders specified a location, are used in this study.

Figure 2: A heatmap showing the geographic distribution of user-labeled videos. The gradient from light green (less) to dark red (more) indicates the relative number of videos geocoded at any location.
III. Feature Extraction

As stated above, there are numerous signals that can be used to infer the geolocation of a video. Not only is the content of the video useful, but any information associated with it may allude to the location at which the video was recorded. In this section, the various information sources and feature representations are discussed.

A. Audio-Visual Features

Local Interest Points: Visual features are a powerful indicator of the geographic location depicted in a video. Through analyzing the distribution of colors in a video one can easily distinguish between, for example, a beach and a ski slope. However, distinguishing some locations may require very specific local visual features: two visually similar places may be distinguished, for example, by the presence or absence of a snow-capped mountain in the distance. However, computing, storing, and matching densely sampled, complex visual features is prohibitively expensive for a large corpus of videos. For this study, local visual features are sampled at sparse interest points using a Laplacian of Gaussian feature extractor. At each such interest point, a local descriptor is computed using Gabor wavelet responses at different orientations, spatial scales, and spatial offsets from the interest point [11]. A hierarchical k-means algorithm, proposed by Nistér and Stewénius [12], is used to build 20,000-word codebooks from the local features. This sparse feature is extracted every 0.9 second throughout each video.

Hue-Saturation Histogram: A hue-saturation time series is constructed from an 8×8 hue-saturation histogram for each video frame. A 8-level 1D Haar wavelet decomposition is applied to each time series. Finally, mean, variance, and extrema at each scale are used as features.

Motion and Shot Detection: Approximate global motion was estimated by computing the cosine distance between consecutive frames. The values are mapped to a temporal sequence by computing the mean distance on consecutive temporal windows of 0.3 s. The 8-level 1D Haar wavelet decomposition is applied to the sequence and features are computed as for the Hue-Saturation histogram. A shot boundary detection algorithm similar to Ren et al. [13] is employed to determine the frequency of shot boundaries within each video. The presence of shot boundaries is detected within consecutive temporal windows of 0.3 s. The result is mapped to a temporal sequence of boolean values corresponding to the absence or presence of a boundary. Features are computed as in the previous cases.

Image-Based Features: A number of image-based features, described in Rowley et al. [14] are extracted for each frame of video. These features, originally developed to detect adult content, encode the presence of people, skin, faces and the proportion of the video frame they cover. They are encoded as a time series for each video using a wavelet decomposition proposed by Gargi and Yagnik [15].

Audio Features: A 32-bin audio spectrogram and the volume are computed for each audio frame. The result is averaged to achieve a sampling period of 0.3 s. The value of each bin and the volume are treated as temporal sequences and processed as in the previous cases.

B. Audio-Visual Categorization

The audio-visual features described in the preceding section form for each video a feature vector consisting of tens of thousands of binary and continuous features. The dimensionality and complexity of this feature space present a challenge for learning, both computationally and in terms of the generalizability of the resulting classifier. The audio-visual features are thus mapped to an audio-visual categorization. The categorization is performed according to the methodology described by Toderici et al. [16]. In addition to a reduced dimensionality, this strategy garners the advantage that the categorical mapping offers some robustness to fitting very specific combinations of audiovisual features, such as video game cut scenes, news broadcasts, music, logos, etc.

The 89 most reliable categories, as determined in [16], were encoded as dense features for each video. A separate feature was encoded for each category where the value was a real-valued confidence measure of how relevant the category is to the video.

C. Landmark Detection

One of the most revealing clues about a location is the observation of a unique landmark. Observing the Eiffel Tower in a video, for example, is an almost certain indication that the video was recorded in Paris, France.

The web-scale landmark detection engine of Zheng et al. [10] is used to detect landmarks in video. However, due to computational limitations only one frame per video is queried for landmarks. This frame was selected near the temporal center of the video, containing low motion, a good candidate for a representative thumbnail of the video.

D. Text-based Features

Text metadata is often very informative about a video’s geolocation. A video description may plainly state it: “This is a video of Zurich, Switzerland.” However, the association to a geographic location may also be inferred from more subtle cues: a reference to a vineyard may suggest that the video was filmed in Napa, California or Bordeaux, France. Determining which words and combinations thereof are good indicators of a location is a challenging task due to the size and complexity of the vocabulary used on the internet. Popular strategies such as stemming and tf-idf may help narrow down the realm of potentially useful words, but still the immense size of a bag-of-words feature vector resulting from such strategies makes learning intractable.
In this study words are clustered using Noisy-Or Bayesian Networks to infer highly likely clusters given a set of words [17]. The lexicon of the model consists of common words, bigrams and phrases. The model parses an inputted set of text and, for each match to the lexicon, activates the corresponding terminal of the Bayesian Network. Both the most likely clusters given these terminals and then the most likely terminals given the clusters are inferred. Finally, a taxonomic classifier (a linear support vector machine) is used to map the cluster activations to a predefined set of categories. The categorization is comparable to that of the Open Directory Project (http://www.dmoz.org/), but only a subset of popular and geographic categories are used. This taxonomic categorization is similar to that of Song et al. [18] but with only textual input. For each video, the title, description, and user-inputted keywords are used to derive these text-based features.

E. Uploader-derived Features

The location at which a video was uploaded and the location at which it was recorded are closely related and often the same. However, some of the most geographically interesting videos are those which are uploaded at a different location from where they were taken. Tourist videos are often recorded at an interesting tourist site and then uploaded in a person’s home city. More complex relationships exist between the upload location and the geolocation that reflect common travel patterns. Unfortunately, a timestamp does not exist for most videos, so one cannot rule out locations due to impossible travel times and distances such as is done to locate photos by Kalogerakis et al. [6]. Also, there is no explicit signal that indicates the location from which a video was uploaded. Instead, publicly available user-supplied information is used to coarsely deduce this. In this work, two signals are used to extract features representing the uploader’s location. One is the ‘hometown’ and country specified by the user (however, this, particularly the hometown, is often missing), which are treated as binary features. This information is publicly available on Anonymized.com, and thus used, only if the user consented to publicizing it. It is generally only inputted once, and thus can be misleading if the user is in a location other than that specified (e.g. traveling). The second signal is the time of the upload, which is also publicly available. The upload hour, in the universal time zone (UTC) using the 24 hour clock, and month were encoded as real-valued features to coarsely capture upload trends for locations.

IV. METHODOLOGY

Training classifiers to distinguish thousands of classes using millions of training examples and hundreds of thousands of features is a challenging task. The size of the feature space makes learning intractable using many standard machine learning approaches. Moreover, many of the mixed binary and continuous features apply only to a small subset of classes; observing the text feature vineyard is not particularly informative for New York City, but it is for Bordeaux. Thus, feature selection plays an important role in this work to make learning tractable, prevent overfitting and reduce the memory footprint of the resulting classifiers.

In this study, decision stumps are used as weak learners. Each decision stump operates as a simple single-feature threshold-based classifier, returning true only if the feature is equal to or above the threshold. Stumps could have either positive or negative thresholds and no limit was placed on the number of stumps per feature. The Adaboost learning algorithm [19] is used to train a weighted combination of these decision stump classifiers. During each stage of Adaboost learning, all possible thresholds on all the features are evaluated and the one that improves the weighted training error the most is selected and stored as a decision stump. Parallelizing this procedure such that each feature is evaluated using a separate thread is straightforward and facilitates fast training in a very large feature space. The resulting classifier, although based on a weighted combination of simple rules, can create a very complex decision boundary. Furthermore, provided that the number of stumps is less than the total number of features, the decision stump learning process inherently performs feature selection, and the computational time and memory complexity of the resulting classifier is proportional to the number of stumps rather than the number of features. From the set of 20 million geotagged videos, approximately 15% are withheld as a test set, with the rest used for training.

Videos were geographically organized according to political boundaries. While latitude-longitude pairs and a multi-scale grid of geographic coordinates (similar to [6]) were considered, political boundaries were chosen for a number of reasons. They offer a natural multi-resolution organization of videos (rather than binning into cells with arbitrary boundaries) and facilitate a more naturally interpretable prediction. Certainly, the ability to pinpoint a video to geographic coordinates would be ideal, but regression in the latitude-longitude space is extremely challenging (especially when the multitude of indoor and geographically irrelevant videos in Anonymized.com is considered) and the size and dimensionality of the training data prohibits a practical nearest neighbors based solution. Thus, for each video, the user-inputted geographic coordinates or text label were mapped to a postal address. From this address a hierarchy of political divisions of at most four levels was extracted. This hierarchy corresponds to the four largest political divisions in the address with a maximum resolution at the city level. These levels will be referred to as the locality (city), administrative subdivision (e.g. U.S. counties), administrative division (e.g. U.S. states or U.K. provinces) and country. Each distinct geographic location in each level of the hierarchy for which there are sufficient labeled videos is considered a geographic
label. Perhaps unsurprisingly, the distribution of total videos per label resembles a power law. A separate binary Adaboost classifier is trained for each geographic label – thus separate classifiers exist for the labels “New York, New York, United States,” “New York, United States” and “United States” with possibly overlapping training data.

A classifier is trained for each location, considered a class, according to the following strategy: the training set uses at most five videos per uploader. All positive videos in the training set are used and 250,000 videos are randomly sampled from all videos that don’t belong in the positive set. During each iteration of training, the Adaboost algorithm selects a threshold on the feature that maximally distinguishes videos (according to the Adaboost reweighted training set) of that geographic location from all other locations. The resulting decision stump is then weighted such that the combined classifier maximizes the classification margin between positive and negative samples.

As Adaboost is well established in the literature, it will not be detailed here. However, some interesting considerations, particularly the tradeoff between classifier complexity and the amount of training data, directed the methodology and feature extraction process. Reyzin and Schapire [20] demonstrated that increasing the complexity of the base classifiers in boosting can increase generalization error despite decreased training error. The generalization error has been shown for any $\theta > 0$ to be at most:

$$Pr[\text{margin}(x, y) \leq \theta] + O\left(\frac{d}{m\theta^2}\right),$$

where $d$ is the Vapnik–Chervonenkis dimension (complexity) of the space of all possible base classifiers, $x$ and $y$ are a training pair input and label respectively, $m$ is the number of training samples and $Pr$ denotes empirical probability on the training sample [19]. For a decision stump based classifier, $d$ increases proportionally to the feature dimensionality and significantly more for dense continuous features than binary ones (a single decision stump can separate binary feature values). Initial attempts to classify based on the raw audio-visual features resulted in poorer generalization error than using the audio-visual categories and this is hypothesized as the explanation why. The smaller-dimensional categorical audio-visual features facilitated the use of more training data, increasing $m$, while decreasing $d$.

A desirable property of the Adaboost algorithm is that the predictive score, or margin, is a real-valued confidence measure. Thus, by applying a threshold on this confidence, one can tune the true and false positive rates of the resulting classifier. Furthermore, classifications based on a weighted combination of decision stumps are easily interpretable. This observation presents some interesting avenues for further research, as the distinguishing features characterizing videos from various places reveal some interesting anthropological and cultural insights.

V. Experiments and Results

A. Landmark Detection

The landmark detector of Zheng et al [10] was employed to detect landmarks in the entire set. Of the 20 million videos, the landmark detector returned approximately 50,000 positive matches for landmarks from over 3,000 locations in 130 countries ranging from Albania to Zimbabwe. Although the geographic precision of the detector was observed to be near-perfect, only 0.25% of videos had matched landmarks. Thus for geolocation this strategy is useful in domains where high precision, high quality geolocation takes precedence over the fraction of videos that can be geolocated.

Of the few failure cases of the landmark detection strategy, some notable ones were moving landmarks, such as specific airplanes photographed at various airports, and videos containing text, which were occasionally mapped to frequently photographed signs or gravestones.

B. Large Scale Video Geolocation

The classification methodology described in Section IV was employed to determine the geographic location of videos on a larger scale.

Approximately 7,100 locations were considered for classification, corresponding to the set of all locations for which there were at least 2000 labeled videos in the training set. A classifier was trained for each of these 7000 locations. The adaboost training was stopped when the classifier reached a specific number of stumps, corresponding to 10% of the size of the training set (with a maximum of 1800 stumps).

A validation set, created by withholding 15% of the training videos from the training procedure was used to determine the equal error rate of each classifier. If this was over 20%, the classifier was discarded. Otherwise it was added to the final set of classifiers. This strategy resulted in classifiers for a total of 56 countries, 458 administrative divisions, 1087 administrative subdivisions and 624 cities.

Using this set of 2225 classifiers, the full test set (distinct from the validation set) was used to determine the true and false positive rates of the resulting classifiers. Specifically, for each classifier, each of the approximately 3 million test videos were classified as being from that location or not. A threshold on the classifier’s confidence was tuned using equal steps to produce Receiver Operating Characteristic (ROC) curves for each classifier. The true and false positive rates for each threshold were averaged over all 2225 classifiers to show the overall average ROC curve in Figure 4 (top) for different subsets of features (on the same set of locations). The ROC curves are shown for various locations in Figure 4 (bottom), using all features.

Figure 4 displays the 45 highest confidence predictions for three interesting locations. Table I shows for these locations the top ten features upon which the cumulative absolute weight assigned to the decision stumps was the highest.
For each video in the test set, a location was predicted as that corresponding to the highest confidence classifier returning a positive classification. The geographic coordinates corresponding to the location were determined as the geographic center of the predicted location. The geolocation precision of the 1.16 million predictions obtained in this way is depicted in Figure 5. Of these predictions, only 7.45% were found to have predicted a location in the wrong country.

Another experiment was conducted to estimate the geographic coverage of the classifiers on unbiased (unlabeled) videos using a confidence threshold set at a true positive rate of 20% and a false positive rate of 0.024%. 8 million videos were sampled from the corpus of videos without ground truth geographic labels and classified, the geographic distribution of which is displayed in Figure 3.

![Figure 3: The predictive distribution of the video geolocation classifier presented in this study. This heatmap demonstrates the predicted locations of all high confidence predictions in a set of over three million unlabeled videos.](image)

Figure 3: The predictive distribution of the video geolocation classifier presented in this study. This heatmap demonstrates the predicted locations of all high confidence predictions in a set of over three million unlabeled videos.

VI. DISCUSSION

The results presented in Section V are very promising and clearly demonstrate that video geolocation at high accuracy is possible for numerous videos and locations. Conversely, certain locations are very difficult to classify and for many videos a geographic location is simply not relevant. Consider, for example, the numerous videos which are recorded indoors. Anonymized.com has many videos consisting of closeup faces, party videos, indoor concerts, pet videos and more. These videos will be very challenging to geolocate with confidence and precision. Through making only high confidence predictions, the methodology presented can ignore these videos, preferring not to make a prediction if there is uncertainty.

As the locations are derived from a political hierarchy, it may seem natural to follow a hierarchical classification strategy by first classifying videos by country, then by administrative area, and so on. However, it was often found in this work that classifying by country was more challenging than by smaller divisions. After filtering out classifiers with a too-low equal error rate, only 56 country classifiers remained. However, from the set of 2225 remaining classifiers there are classifiers from 158 countries, including the 56 countries and an additional 102 countries for which there are classifiers for political divisions within those countries. In larger countries with more varied areas, visual features may be less useful and specific text strings (such as “ski”) less meaningful.

Although there is no literature on video geolocation to directly compare the results presented in this study to, it may be appropriate to consider the work of Hays and Efros [5]. Comparing Figure 5 to the equivalent figure (Figure 6) of [5] shows that this methodology can geolocate videos at a significantly higher geographic precision than [5] can geolocate images using a mean-shift or nearest neighbors approach. 25% of their images are predicted at up to the country resolution while 25% of the videos in this study can be predicted at the city resolution. However, there are many differences in the data sets, feature representations and methodologies of the two approaches that limit the conclusiveness of this comparison.

Analysis of the features upon which the classifiers determine location reveals very interesting trends in how videos
Figure 5: The geolocation precision of the location predictions in kilometers. For each video in the test set, a location was predicted as that corresponding to the highest confidence classifier returning a positive classification. The geographic coordinates corresponding to the location were determined as the center point of the predicted location. Thus a prediction of USA for a video recorded in New York City would have a geolocation error consisting of the distance from the center of the USA to the center of New York City.

Table I: The top ten weighted features for each location. This table shows for some example locations the features upon which decision stumps received the most weight in the Adaboost classifier. These can be considered as the features upon which rules of thumb contribute the most to the confidence about whether a video was recorded in a location.

<table>
<thead>
<tr>
<th>Location</th>
<th>Top 10 Weighted Features</th>
</tr>
</thead>
</table>

Figure 6: The top confidence predictions for various example locations from the test set. The video thumbnails for the 45 top confidence predictions for each location are shown. Captured this and prefers videos that visually depict soccer matches and are uploaded during the soccer season by an uploader from Manchester. Le Mans, France is the setting of the yearly “24 Hours of Le Mans” sports car race. The classifier in this case gives a high confidence score for videos that depict car racing, have text related to car racing, and are uploaded during a particular timeframe of a couple of summer months in France. Interestingly, the classifier applied a large negative weight to uploaders from Brazil.
This initially was considered erroneous until, upon further inspection, it was discovered that in fact a portion of the Le Mans race series was held in Brazil. The classifier learned that many false positives depicted car racing and mentioned “Le Mans” in the metadata, but the uploader specified they were from Brazil. The classifier downweighted the corresponding feature accordingly.

Figure 4 showed that the combination of all features outperforms any subset of features. However, one may consider which types of features are most informative for the task of geolocation. Figure 7 shows the total weight applied to each feature type by the Adaboost classifiers. Approximately 40% of the total weight was distributed to the visual categories and 30% each to the other feature types. Of the text features, the text terminals were weighted highest, followed by the categories. Thus, all the features are important, but visual features were weighted highest. These may be more useful than the other features to help filter out videos which are difficult to classify (i.e. indoor videos, parties, etc.).

Remarkably, most locations have the upload hour as their highest weighted feature. This fact likely reflects uploading trends that coarsely reveal the location of the uploader. Videos from much of the world can be eliminated if they have unlikely upload times (such as early in the morning) for a particular location. This combined with the uploader’s declared location likely helps determine whether the video was in fact uploaded in this location.

Analysis of the various classifiers has revealed very interesting structure and trends for certain locations. Unfortunately, no subset of features is shared by the majority of locations, which prohibits storing and computing a more compact feature representation. For brevity, only a small subset of interesting trends discovered through this study are demonstrated here. Further analysis will be a very interesting avenue for future work.

REFERENCES


