Audio Content based Geotagging in Multimedia

Anurag Kumar  
Carnegie Mellon University  
Pittsburgh, PA, USA  
alhu@andrew.cmu.edu

Benjamin Elizalde  
Carnegie Mellon University  
Moffett Field, CA, USA  
bmartin1@andrew.cmu.edu

Bhiksha Raj  
Carnegie Mellon University  
Pittsburgh, PA, USA  
bhiksha@cs.cmu.edu

ABSTRACT

In this paper we propose methods to extract geographically relevant information in a multimedia recording using only its audio component. Our method primarily is based on the fact that urban acoustic environment consists of a variety of sounds. Hence, location information can be inferred from the composition of sound events/classes present in the audio. More specifically, we adopt matrix factorization techniques to obtain semantic content of recording in terms of different sound classes. These semantic information are then combined to identify the location of recording.

CCS Concepts

• Information systems → Speech / audio search; Information extraction; Video search;

Keywords

Location Identification, Geotagging, Matrix Factorization, Kernel Fusion

1. INTRODUCTION

Extracting information from multimedia recordings has received lot of attention due to the growing multimedia content on the web. A particularly interesting problem is extraction of information related to geographical locations. The process of providing information about geographical identity is usually termed as Geotagging [11] and is gaining importance due its role in several applications. It is useful not only in location based services and recommender systems [1] [2][12] but also in general cataloguing, organization, search and retrieval of multimedia content on the web. Location specific information also allows a user to put his/her multimedia content into a social context, since it is human nature to associate with geographical identity of any material. A nice survey on different aspects of geotagging in multimedia is provided in [11].

In all of the contexts mentioned above the central point is the availability of geographical information about the multimedia. Although, there are applications which allows users to add geographical information in their photos and videos, a larger portion of multimedia content on the web is without any geographical identity. In these cases geotags needs to be inferred from the multimedia content and the associated metadata. This problem of geotagging or location identification also features as the Placing Tasks in yearly MediaEval tasks. The goal of Placing Tasks in MediaEval is to develop systems which can predict places in videos based on different modalities of multimedia such as images, audio, text etc. An important aspect of location prediction systems is the granularity at which location needs to be predicted. The Placing Task recognizes a wide range of location hierarchy, starting from neighbourhoods and going upto continents. In this work we are particularly interested in obtaining city-level geographical tags which is clearly one of the most important level of location specification for any data.

Most of the current works on geotagging focus on using visual/image component of multimedia and the associated text in the multimedia ([10] [11][13][8] to cite a few). The audio component has been largely ignored and there is little work on predicting location based on audio content of the multimedia. However, authors in [6] argue that there are cases where audio content might be extremely helpful in identifying location. For example, speech based cues can aid in recognizing location. Moreover, factors such as urban soundscapes and locations acoustic environment can also help in location identification. A few works such as [10] [17] did tried to exploit audio cues for geotagging in videos. However, the approaches proposed have been simplistic relying mainly on basic acoustic features. The general schema is to either directly use basic acoustic features such as Mel-Cepstra Coefficient (MFCC) or to obtain audio-clip level features such GMM-Supervectors or Bag Of Audio Words histograms and then build classifiers on these features.

In this work we show that geotagging using only audio component of multimedia can be done with reasonably good success rate. Our primary assertion is that the semantic content of an audio recording in terms of different sound events can help in predicting locations. We argue that soundtracks of different cities are composed of a set of sound events. If we can somehow capture the composition of audio in terms of these sound events then they can be used to train machine learning algorithms for geotagging purposes. We start with a set of base sound events or classes and then use methods based on matrix factorization to find the composition of soundtracks in terms of these sound events. Once the weights corresponding to each base sound class have been obtained, we build higher level feature using these weights which are further used to obtain kernels representations. The kernels corresponding to each base sound are then combined to finally train Support Vector Machines for predicting location identification of the recording.

The rest of the paper is organized as follows. In Section 2 we describe our proposed framework for audio based geotagging. In Section 3 we present our experiments and results.
In Section 4 we discuss scalability of our proposed method and also give concluding remarks.

2. AUDIO BASED GEOTAGGING

Audio based geotagging in multimedia can be performed by exploiting audio content in several ways. One can possibly try to use automatic speech recognition (ASR) to exploit the speech information present in audio. For example, speech might contain words or sentences which uniquely identifies a place. I am near Eiffel Tower clearly gives away the location as Paris with high probability irrespective of presence or absence of any other cues. Other details such as language used, mention of landmarks etc. in speech can also help in audio based geotagging. However, in this work we take a more generic approach where we try to capture semantic content of audio through occurrence of different meaningful sound events and scenes in the recording. We argue that it should be possible to train machines to capture identity of a location by capturing the composition of audio recordings in terms of human recognizable sound events. This idea can be related to and is in fact backed by urban soundscapes works [3, 12]. Based on this idea of location identification through semantic content of audio, we try to answer two important questions. First how to mathematically capture the composition of audio recordings and Second how to use the information about semantic content of the recording for training classifiers which can predict identity of location. We provide our answers for each of these questions one by one.

Let $E = \{E_1, E_2, \ldots E_l\}$ be the set of sound events which we want to capture in audio recordings. $E_1$ to $E_l$ are different sound events or classes. We assume that each of these sound classes can be characterized by a basis matrix $M_l$. For a given sound event $E_l$ the column vectors of its basis matrix $M_l$ essentially spans the space of sound event $E_l$. Mathematically, this span is in space of some acoustic feature (e.g MFCC) used to characterize audio recordings and over which the basis matrices have been learned. How we obtain $M_l$ is discussed later. Any given soundtrack or audio recording is then decomposed with respect to a sound event $E_l$ as

$$X \approx M_l W_l^T$$

(1)

where $X$ is a $d \times n$ dimensional representation of the audio recording using acoustic features such as MFCC. For MFCC, this implies each column of $X$ is $d$ dimensional mel-frequency cepstral coefficients and $n$ is the total number of frames in the audio recording. The sound basis matrices $M_l$ are $d \times k$ dimensional where $k$ represents the number of basis vectors in $M_l$. In principle $k$ can vary with each sound class, however, for sake of convenience we assume it is same for all $E_l$, $l = 1 \text{ to } L$.

The idea behind Eq. 1 is to obtain $W_l$ which captures how the sound class $E_l$ is present in the recording. The weight matrix $W_l$ captures the presence of each sound event through out the duration of the recording. Obtaining $W_l$ for each $l$ provides us information about the structural composition of the audio in terms of sound classes in $E$. Hence, these $W_l$ can be used for differentiating locations. Now, the problem boils down to learning $M_l$ for each $E_l$ and then using it to compute $W_l$ for any given recording.

2.1 Obtaining $M_l$ and $W_l$

Let us assume that for a given sound class $E_l$ we have a collection of $N_l$ audio recordings belonging to class $E_l$ only. We parametrize each of these recordings through some acoustic features. In this work we use MFCC features augmented by delta and acceleration coefficients (denoted by MFCA) as basic acoustic features. These acoustic features are represented by $d \times n_i$ dimensional matrix $X_{E_l}$ for the $i^{th}$ recording. $d$ is dimensionality of acoustic features and each column represents features for a frame. The basic features of all recordings are collected into one large matrix $X_{E_l} = [X_{E_l1}, \ldots X_{E_lN_l}]$ to get a large collective sample of acoustic features for sound class $E_l$. Clearly, $X_{E_l}$ has $d$ rows and let $T$ be the number of columns in this matrix.

To obtain the basis matrix $M_l$ for $E_l$ we employ matrix factorization techniques. More specifically, we use Non-Negative matrix factorization (NMF) like method proposed in [7]. [4] proposed two matrix factorization methods named semi-NMF and convex-NMF which are like NMF but do not require the matrix data to be non-negative. This is important in our case, since employing classical NMF [9] algorithm would require our basic acoustic feature to be non-negative. This can be highly restrictive given the challenging task at hand. Of the two methods semi-NMF and convex-NMF, semi-NMF yielded better results in our experiments and hence due to space constraints we present description and results for only semi-NMF in this paper.

semi-NMF considers factorization of a matrix, $X_{E_l}$ as $X_{E_l} \approx M_l W_l^T$. For factorization number of basis vectors $k$ in $M_l$ is fixed to a value less than $\text{min}(d, T)$. semi-NMF does not impose any restriction on $M_l$ that is its element can have any sign. The weight matrix $W$ on the other hand is restricted to be non-negative. The objective is to minimize $||X_{E_l} - M_l W_l^T||^2$. Assuming that $M_l$ and $W$ have been initialized, [7] gave the following iterative update rules for factorization. In each step of iteration,

- Fix $W_l$ update $M_l$ as,
  $$M_l = X_{E_l} W_l (W_l W_l^T)^{-1}$$
  (2)

- Fix $M_l$, update $W_l$, $W_{rs} = W_{rs} \sqrt{(X_{E_l}^T M_l)_{rs} + (W_l^T M_l)_{rs} - |(X_{E_l}^T M_l)_{rs}|^2 / (|W_l^T M_l|_{rs})^2)}$
  (3)

The process is iterated till error drops below certain tolerance. The + and − sign represents positive and negative parts of a matrix obtained as $Z_{rs} = (|Z_{rs}| + Z_{rs})/2$ and $Z_{rs} = (|Z_{rs}| - Z_{rs})/2$. Theoretical guarantees on convergence of semi-NMF and other interesting properties such as invariance with respect to scaling can be found in original paper. One interesting aspect of semi-NMF described by authors is its analysis in terms of K-means clustering algorithm. The objective function $||X - MW^T||^2$ can be related to K-Means objective function with $M_l$ having the $k$ cluster centers. Hence, the basis matrix $M_l$ also represents centers of a group of clusters. We actually exploit this interpretation in the next phase of our approach. The initialization of $M_l$ and $W_l$ is done per the procedure described in [7].

Once $M_l$ have been learned for each $E_l$, we can easily obtain $W_l$ for any given audio recording $X$ by fixing $M_l$ and then applying Eq 3 for $X$ for several iterations. For a given $X$, $W_l$ contains information about $E_l$ in $X$. With K-Means interpretation of semi-NMF, the non-negative weight matrix $W_l$ can be interpreted as containing soft assignment posteriori to each cluster for all frames in $X$. 

2.2 Discriminative Learning using W_l

We treat the problem of location prediction as a retrieval problem where we want to retrieve most relevant recordings belonging to a certain location (city). Put more formally, we train binary classifiers for each location to retrieve the most relevant recordings belonging to the concerned location. Let us assume that we are concerned with a particular city C and the set \( S = \{ s_i, i = 1 \to N \} \) is the set of available training audio recordings. The labels of the recordings are represented by \( y_i \in \{ -1, 1 \} \) with \( y_i = 1 \) if \( s_i \) belongs to \( C \) and otherwise \( y_i = -1 \). \( X_i \) \((d \times n_i)\) denotes the MFCA representation of \( s_i \). For each \( X_i \), we compute composition matrices \( W_i^l \) obtained with respect to all sound events \( E_l \) in \( E \). \( W_i^l \) captures distribution of sound event \( X_i \) and we propose 2 histogram based representations to characterize this distribution.

2.2.1 Direct characterization of \( W_i \) as posterior

As we mentioned before semi-NMF can be interpreted in terms of K-means clustering. For a given \( E_l \), the learned basis matrix \( M_l \) can be interpreted as matrix containing cluster centers. The weight matrix \( E_l \) is used to train SVM for prediction. Since exponential \( \chi^2 \) kernel SVM are known to work well with histogram representations \( [20][21] \), we use kernels of the form \( K_l(\vec{r}_i, \vec{r}_j) = \exp(-D(\vec{r}_i, \vec{r}_j)/\gamma) \) where \( D(\vec{r}_i, \vec{r}_j) \) is \( \chi^2 \) distance between \( \vec{r}_i \) and \( \vec{r}_j \). \( \gamma \) is set as the average of all pair wise distance. Once we have all \( K_l \), we use two simple kernel fusion methods; average kernel fusion where \( K_{\text{av}} = \frac{1}{L} \sum_{l=1}^{L} K_l(\cdot, \cdot) \) and product kernel fusion where \( K_{\text{pr}} = \prod_{l=1}^{L} K_l(\cdot, \cdot) \). The final kernel representation \( K_{\text{av}} \) or \( K_{\text{pr}} \) is used to train SVM for prediction.

3. EXPERIMENTS AND RESULTS

We evaluate our proposed method on a subset of videos from MediaEval Placing Task [10]. The dataset contains a total of 1079 Flickr videos with 540 videos in the training set and 539 in the testing set. We work with only audio of each video and we will alternatively refer to these videos as audio recordings. The maximum duration of recordings is 90 seconds. The videos of the recording belong to 18 different cities with several cities having very few examples in training as well as testing set such as just 3 for Bankok or 5 for Beijing. We selected 10 cities for evaluation for which training as well as test set contains at least 11 examples. These 10 cities are Berlin (B), Chicago (C), London (L), Los Angeles (LA), Paris (P), Rio (R), San Francisco (SF), Seoul(SE), Sydney (SY) and Tokyo (T). As stated before the basic acoustic feature used are MFCC features augmented by delta and acceleration coefficients. 20 dimensional MFCCs are extracted for each audio recording over a window of 30 ms with 50% overlap. Hence, basic acoustic features for audio recordings are 60 dimensional and referred to as MFCA features.

For our proposed method we need a set of sound classes \( E_l \). Studies on Urban soundscapes have tried to categorize the urban acoustic environments [3][13][14][15] came up with a refined taxonomy of urban sounds and also created a dataset, UrbanSounds8k, for urban sound events. This dataset contains 8732 audio recordings spread over 10 different sound events including automotive noise, outdoor noise, street music. We use these 10 sound classes as \( E_l \) and we obtain the basis matrices \( M_l \) for each \( E_l \) using the examples of these sound events provided in the UrbanSounds8k dataset. The number of basis vectors for all \( M_l \) is same and fixed to either 20 or 40. We present results for both cases. Finally, in the classifier training stage; SVMs are trained using the fused kernel \( K_{\text{av}} \) or \( K_{\text{pr}} \) as described in Section 2.2.2. Here the slack parameter \( C \) in SVM formulation is set by performing 5 fold cross validation over the training set.

We compare our proposed method with bag of audio words (BoAW) representation built over MFCA acoustic features. This method builds bag of words features directly over basic acoustic features and is known to work well. The first step in this method is to train a GMM \( G^\text{av} \) with \( G^\text{av} \) components over MFCA features where each Gaussian represents an audio word. Then for each audio recording clip level histogram features are obtained using the GMM posteriors for each Vector Machine (SVM) to finally train classifiers for geotagging purposes. We explain our strategy here in terms of \( \vec{r}_i \) and \( \vec{v}_i \) the same steps are followed. For each \( l \), we obtain separate kernels representation \( K_l \) using \( \vec{r}_i \) for all \( X_i \). Since exponential \( \chi^2 \) kernel SVM are known to work well with histogram representations \( [20][21] \), we use kernels of the form \( K_l(\vec{r}_i, \vec{r}_j) = \exp(-D(\vec{r}_i, \vec{r}_j)/\gamma) \) where \( D(\vec{r}_i, \vec{r}_j) \) is \( \chi^2 \) distance between \( \vec{r}_i \) and \( \vec{r}_j \). \( \gamma \) is set as the average of all pair wise distance. Once we have all \( K_l \), we use two simple kernel fusion methods; average kernel fusion where \( K_{\text{av}} = \frac{1}{L} \sum_{l=1}^{L} K_l(\cdot, \cdot) \) and product kernel fusion where \( K_{\text{pr}} = \prod_{l=1}^{L} K_l(\cdot, \cdot) \). The final kernel representation \( K_{\text{av}} \) or \( K_{\text{pr}} \) is used to train SVM for prediction.
frame in the clip. The computation is similar to Eq.5 except that the process is done over MFCA features. We will use $\vec{b}$ to denote these $G^b$ dimensional bag of audio words features. Exponential $\chi^2$ kernel SVMs are again used to train the classifiers. In this case, the parameter $\gamma$ in the kernel along with $C$ is optimized by cross validation over the training set.

As stated before we formulate the geotagging problem as retrieval problem where the goal is to retrieve most relevant audios for a city. We use Average Precision (AP) to measure performance for each city and Mean Average Precision (MAP) over all cities as the overall metric. Due to space constraints we are not able to show AP results in every case and will only present overall metric MAP.

Table 1 shows MAP results for bag of audio word method (top 2 rows) and our proposed method (bottom 3 rows) using $\vec{h}$ features described in Section 2.2.1. For baseline method we experimented with 4 different component size $G^b$ for GMM $\vec{h}^b$. $k$ represents the number of basis vectors in each $M_i$. $K^h_b$ represents the average kernel fusion and $K^b_h$ product kernel fusion. First, we observe that our proposed method outperforms the bag of audio words plus $\chi^2$ kernel SVM by a significant margin. For BoAW, $G^b = 256$ gives highest MAP of 0.478 but MAP saturates with increasing $G^b$ and hence, any significant improvement in MAP by further increasing $G^b$ is not expected. Our proposed method with $k = 40$ and product kernel fusion gives 0.563 MAP, an absolute improvement of 8.5% in MAP when compared to BoAW with $G^b = 256$. MAP in other cases for our proposed method are also better than best MAP for BoAW. We also note that for $\vec{h}$ features, product kernel fusion of different sound class kernels performs better than average kernel fusion. Also, for $\vec{h}$, $k = 40$ is better than $k = 20$.

Table 2 shows results for our $\vec{v}$ features in Section 2.2.2 which uses GMM based characterization of composition matrices $W_i$. We experimented with 4 different values of GMM component size $G^i$. Once again we observe that overall this framework works better than BoAW with exponential $\chi^2$ kernel. We can also observe that if we fix the GMM size as same for BoAW and $\vec{v}$ that is $G^i = G^b = G$, then $\vec{v}$ outperforms BoAW in all cases; upto 9% in absolute terms for $k = 20$ and $G = 32$. This shows that the composition matrices $W_i$ are actually capturing semantic information from the audio and these semantic information when combined helps in location identification. If we compare $\vec{v}$ and $\vec{h}$ methods then overall $\vec{h}$ seems to give better results. This is worth noting since it suggests that $W_i$ on its own are extremely meaningful and sufficient. Another interesting observation is that for $\vec{v}$ average kernel fusion is better than product kernel fusion.

Figure 1 shows city wise results for the three methods. For each method the shown AP correspond to the case of best MAP. This implies GMM component size in both BoAW and $\vec{v}$ is 256 that is $G^i = G^b = 256$; for $\vec{h}$, $k = 40$ and product kernel fusion; for $\vec{v}$, $k = 20$ and average kernel fusion. For convenience, city names have been denoted by indices used in the beginning paragraph of this section. Figure 1 also shows MAP values in the extreme right. Figure 1 shows that cities such as Rio(R), San Francisco(SF), Seoul(SE) are much easier to identify and on the other hand cities like Sydney(SY), Berlin(B) and Tokyo(T) are harder to identify.

4. DISCUSSIONS AND CONCLUSIONS

We presented methods for geotagging using audio content in multimedia recordings. We proposed that if the semantic content of the audio can be captured in terms of different sound events which occur in our environment, then these semantic information can be used for location identification purposes. It is expected that larger the number of sound classes in $E$ the more distinguishing elements we can expect to obtain and the better it is for geotagging. Hence, it is desirable that any framework working under this idea should be scalable in terms of number of sounds in $E$. In our proposed framework the process of learning basis matrices $M_i$ are independent of each other and can be easily parallelized. Similarly, obtaining composition weight matrices $W_i^\lambda$ can also be computed in parallel for each $E_i$ and so do the features $\vec{h}_i$ (or $\vec{v}_i$) and kernel matrices. One can also easily add any new sound class to an existing system if required. Hence, our method can be easily scaled in terms of number of sound events in $E$.

Even with 10 sound events from urban sound taxonomy we obtained reasonably good performance. Our proposed framework outperformed known bag of audio word method by a significant margin. Currently, we used simple kernel fusion methods to combine event specific kernels. One can potentially use established methods such as multiple kernel learning at this step. This might lead to further improvement in results. One can also look into other methods for obtaining basis matrices for sound events. A more comprehensive analysis on a larger dataset with larger number of cities can through more light on the effectiveness of the proposed method. However, this work does give sufficient evidence towards success of audio content based geotagging in multimedia.

| $G^b$ | 32 | 64 | 128 | 256 |
|-------|----|----|-----|-----|
| MAP | 0.362 | 0.429 | 0.461 | 0.478 |

| Kernel | Avg Ker. ($K^h_b$) | Prod. Ker. ($K^b_h$) |
|--------|-----------------|-----------------|
| $k$ | 20 | 40 | 20 | 40 |
| MAP | 0.454 | 0.500 | 0.520 | 0.563 |

| $G^i$ | $k$ | Avg Ker. ($K^i_\gamma$) | Prod. Ker. ($K^i_P$) |
|-------|-----|-----------------|-----------------|
| 32 | 20 | 0.454 | 0.427 | 0.448 | 0.417 |
| 64 | 40 | 0.482 | 0.466 | 0.432 | 0.424 |
| 128 | 256 | 0.510 | 0.465 | 0.466 | 0.427 |
| 256 | 0.527 | 0.455 | 0.471 | 0.441 |

Figure 1: AP for Cities (MAP in right extreme)
5. REFERENCES

[1] J. Bao, Y. Zheng, and M. F. Mokbel. Location-based and preference-aware recommendation using sparse geo-social networking data. In Proceedings of the 20th International Conference on Advances in Geographic Information Systems, pages 199–208. ACM, 2012.

[2] J. Bao, Y. Zheng, D. Wilkie, and M. Mokbel. Recommendations in location-based social networks: a survey. GeoInformatica, 19(3):525–565, 2015.

[3] A. Brown, J. Kang, and T. Gjestland. Towards standardization in soundscape preference assessment. Applied Acoustics, 72(6):387–392, 2011.

[4] L. Cao, S.-F. Chang, N. Codella, C. Cotton, D. Ellis, L. Gong, M. Hill, G. Hua, J. Kender, M. Merler, et al. Ibm research and columbia university trecvid-2011 multimedia event detection (med) system.

[5] J. Choi, H. Lei, V. Ekambaram, P. Kelm, L. Gottlieb, T. Sikora, K. Ramchandran, and G. Friedland. Human vs machine: establishing a human baseline for multimodal location estimation. In Proceedings of the 21st ACM international conference on Multimedia, pages 867–876. ACM, 2013.

[6] J. Choi, B. Thomee, G. Friedland, L. Cao, K. Ni, D. Borth, B. Elizalde, L. Gottlieb, C. Carrano, R. Pearce, et al. The placing task: A large-scale geo-estimation challenge for social-media videos and images. In Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia, pages 27–31. ACM, 2014.

[7] C. Ding, T. Li, and M. I. Jordan. Convex and semi-nonnegative matrix factorizations. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32(1):45–55, 2010.

[8] P. Kelm, S. Schmiedeke, J. Choi, G. Friedland, V. N. Ekambaram, K. Ramchandran, and T. Sikora. A novel fusion method for integrating multiple modalities and knowledge for multimodal location estimation. In Proceedings of the 2nd ACM international workshop on Geotagging and its applications in multimedia, pages 7–12. ACM, 2013.

[9] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In Advances in neural information processing systems, pages 556–562, 2001.

[10] H. Lei, J. Choi, and G. Friedland. Multimodal city-identification on flickr videos using acoustic and textual features. 2012.

[11] J. Luo, D. Joshi, J. Yu, and A. Gallagher. Geotagging in multimedia and computer vision: A survey. Multimedia Tools and Applications, 51(1):187–211, 2011.

[12] A. Majid, L. Chen, G. Chen, H. T. Mirza, I. Hussain, and J. Woodward. A context-aware personalized travel recommendation system based on geotagged social media data mining. International Journal of Geographical Information Science, 27(4):662–684, 2013.

[13] MediaEval. http://www.multimediaeval.org/ 2015.

[14] S. Payne, W. Davies, and M. Adams. Research into the practical and policy applications of soundscape concepts and techniques in urban areas. 2009.

[15] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In Proceedings of the ACM International Conference on Multimedia, pages 1041–1044. ACM, 2014.

[16] R. M. Schafer. The soundscape: Our sonic environment and the tuning of the world. Inner Traditions/Bear & Co, 1993.

[17] X. Sevillaño, X. Valero, and F. Aliás. Audio and video cues for geo-tagging online videos in the absence of metadata. In Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on, pages 1–6. IEEE, 2012.

[18] Y.-C. Song, Y.-D. Zhang, J. Cao, T. Xia, W. Liu, and J.-T. Li. Web video geolocation by geotagged social resources. Multimedia, IEEE Transactions on, 14(2):456–470, 2012.

[19] M. Trevisiol, H. Jégou, J. Delhumeau, and G. Gravier. Retrieving geo-location of videos with a divide & conquer hierarchical multimodal approach. In Proceedings of the 3rd ACM conference on International conference on multimedia retrieval, pages 1–8. ACM, 2013.

[20] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. International journal of computer vision, 73(2):213–238, 2007.