A Visual Approach for Video Geocoding using Bag-of-Scenes

PENATTI, O. A. B. ; LI, L. T. ; ALMEIDA, J. ; TORRES, R. da S.

In: ACM International Conference on Multimedia Retrieval (ICMR), Hong Kong, China, 2012, p. 53:1-53:8.

A Visual Approach for Video Geocoding using Bag-of-Scenes

Otávio A. B. Penatti[1], Lin Tzy Li[1,2], Jurandy Almeida[1], and Ricardo da S. Torres[1]
[1]RECOD Lab, Institute of Computing, University of Campinas,Campinas, SP, Brazil, 13083-852
[2]Telecommunications Res. & Dev. Center, CPqD Foundation, Campinas, SP, Brazil, 13086-902
{penatti,lintzyli,jurandy.almeida,rtorres}@ic.unicamp.br

ABSTRACT

This paper presents a novel approach for video representation, called bag-of-scenes. The proposed method is based on dictionaries of scenes, which provide a high-level representation for videos. Scenes are elements with much more semantic information than local features, specially for geotagging videos using visual content. Thus, each component of the representation model has self-contained semantics and, hence, it can be directly related to a specific place of interest. Experiments were conducted in the context of the MediaEval 2011 Placing Task. The reported results show our strategy compared to those from other participants that used only visual content to accomplish this task. Despite our very simple way to generate the visual dictionary, which has taken photos at random, the results show that our approach presents high accuracy relative to the state-of-the art solutions.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Dictionaries*; H.2.8 [**Database Management**]: Database Applications—*Spatial databases and GIS*; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*

General Terms

Theory, Experimentation

Keywords

video representation, visual words, geotagging, placing task

1. INTRODUCTION

The geographic information is present in people's daily life, thus it is not surprising that there is a huge amount of data in the Web about geographical entities and a great interest in localizing them on maps. That information is often enclosed in digital objects (e.g., documents, image, and videos). Once they are geocoded, one can perform geographical queries. The process of associating a geographic location with photos and videos, which is called geocoding in the Geographic Information Retrieval (GIR) community, is often known as geotagging or georeferencing in the multimedia field [23]. Further, in the Geographic Information System (GIS) area, georeferencing is a term largely used to refer to a given location where something exists, in a physical space, in terms of a coordinate system (i.e., latitude and longitude). Therefore geotagging, georeferencing, or geocoding mean associating the location depicted or referred by a digital object.

Associating a video content with its geographic location has become very popular in many video applications. In order to speed up such a task, geotags can be propagated based on the similarity between video content and context. For this, it is imperative to develop powerful tools for capturing and representing high-level semantics of video data. Identifying and representing semantics of a video content is one of the most important aspects for video analysis, classification, indexing, and retrieval. In most of the current techniques, videos are represented as bag-of-visual-words obtained from dictionaries of local features, like Scale-invariant Feature Transform (SIFT) or Space-time interest points (STIP). Despite the good performance of existing approaches based on this scheme, such a model is based on elements with very few or no semantic information, like corners and edges.

In this paper, we present a novel approach for video representation, named *bag-of-scenes*. The proposed method is based on dictionaries of scenes, which provide a high-level representation for videos. Scenes are elements with much more semantic information than local features, specially for geotagging videos using visual content. The bag-of-scenes video representation works like a *place activation vector* because each scene in the dictionary can be seen as a representative picture from a place. In this way, each component of the feature vector has semantics and, hence, it can be directly related to a specific place of interest.

Our experiments were conducted under the specifications of the Placing Task at MediaEval 2011. The goal of such a task is to automatically assign geographical coordinates (latitude and longitude) to a set of annotated videos [33]. The reported results show the potential of the proposed approach for video geocoding, even considering a simple random selection of scenes to compose the dictionary.

The remainder of this paper is organized as follows. Section 2 introduces some basic concepts and describes related work. Section 3 briefly reviews geocoding approaches from other participants of the Placing Task at MediaEval, which were used as reference in our experiments. Section 4 presents

ICMR '12, June 5-8, Hong Kong, China

our approach and shows how to apply it for representing video data. Section 5 reports the results of our experiments and compares our technique with other methods. Finally, we offer our conclusions and directions for future work in Section 6.

2. BACKGROUND

The technique proposed in this paper is a visual approach for video geocoding using a bag-of-scenes representation. Hence, it is related to the areas of video representation, visual dictionaries, and video geocoding. In the following subsections, we detail the related work of each area separately.

2.1 Video representation

Traditional approaches for video representation are based on global features. In this method, videos are described by the overall distribution of low-level features, such as color, texture, edge, or other visual properties [24, 32, 34].

One limitation of the global methods relies on their sensitiveness to small changes in the spatial distribution. To include spatial information, a keyframe is usually partitioned into either rectangular regions or segments of objects [37,43]. However, the problem of choosing an appropriate scale to compare visual features still persists. Using a more local scale increases the susceptibility of the method in presence of object and camera motions, while a more global scale decreases its sensitivity to changes in the spatial distribution.

More recently, solutions based on local keypoints have emerged as an alternative to overcome the problem of object and camera motion. Keypoints are salient patches with a rich information for describing details and nuances of the scenes. They can be detected using feature detectors, such as difference of Gaussian (DoG) [22] and Hessian-Laplace regions [25]; and depicted by local descriptors, such as SIFT [22] and STIP [16]. A comprehensive review of feature detectors and local descriptors can be found in [26, 27, 38].

There are two different approaches of utilizing those features. Keypoints can be matched directly in the feature space and then the matching patterns [28] or the cardinality of matching pairs [12] can be used to estimate the video similarity. Alternatively, they can be vector-quantized or clustered into a representation model based on bag of words, commonly used in the Information Retrieval area. Many works in the literature have used this video representation, popularly referred to as BoW, for multimedia retrieval and classification [10, 36, 44].

2.2 Visual dictionaries

Visual dictionaries, whose image representations are the so-called *bag-of-words*, have the advantage of preserving the discriminative power of local descriptions while pooling those features into a single feature vector [6].

To compute a bag-of-words (BoW) representation for an image, one must first create a visual dictionary or codebook to describe the image according to visual words. To generate a visual dictionary, local low-level features are first extracted from images. One can use feature detectors or employ dense sampling in a regular grid to obtain the image regions to be described. In the literature, the latter approach has outperformed the former one on classification tasks [39]. Each image region is then described by a local descriptor, with SIFT being the most popular choice [22]. Those feature vectors are thus clustered or randomly sampled in order to obtain the visual words of the dictionary. Although k-means is still a common technique for clustering the feature space, a simple random selection of points generates dictionaries of similar quality, due to the curse of the dimensionality [42]. It is important to highlight that the clustering of the feature space is based solely on patch appearance, and, given the fact that patches were extracted from small punctual regions of images, like corners and edges, they themselves carry no semantics [14].

The created dictionary is used to generate a image representation. This is performed by assigning one or more of the visual words to each point in an image. One can use *hard* or *soft* assignment, with the last being more robust to feature quantization problems [21, 40]. Soft assignment of a point p_i to a dictionary of k words can be formally given by Equation 1 [40]:

$$\alpha_{i,j} = \frac{K_\sigma(D(p_i, w_j))}{\sum_{l=1}^{k} K_\sigma(D(p_i, w_l))}, \quad (1)$$

where j varies from 1 to k, $K_\sigma(x) = \frac{1}{\sqrt{2\pi}\times\sigma} \times exp(-\frac{1}{2}\frac{x^2}{\sigma^2})$, and $D(a, b)$ is the distance between vectors a and b.

After the description of points using the created dictionary, a pooling step is applied. The pooling step generates the final image feature vector by summarizing the assignment values of each point. Popular pooling strategies are *average* and *max* pooling, with an advantage to the last one [6]. Max pooling can be formally defined by Equation 2:

$$h_j = \max_{i \in N} \alpha_{i,j} \quad (2)$$

where N is the number of points in the image and j varies from 1 to k.

Due to the popularity and high effectiveness of the visual dictionary model, the research community is very active in this area. A performance evaluation of different detectors and descriptors has been presented by van de Sande et al. [39]. Several other works have focused on improving the assignment step [5, 21, 40] or evaluating and proposing new pooling strategies [6, 19, 30].

The use of a similar strategy to describe videos seems to be reasonable. Our approach tries to comprise all the advantages of a visual dictionary model by yet including semantics in the visual words.

2.3 Video geocoding

Current solutions for geocoding multimedia material are usually based on textual information [17, 23]. Such a strategy depends on the human intervention to associate textual descriptions with multimedia data. However, there is a lack of objectivity and completeness of those descriptions, since the understanding of the visual content of multimedia data may change according to the experience and the perception of each subject. Other issues are related to lexical and geographical problems in recognizing place names [18]. Those limitations open new venues for the investigation of methods that use image/video content in the geocoding process.

Predicting a location based on a given image was the goal of Hays and Efros [9], whose strategy was to find a probability distribution of images over the globe and base their strategy on that information, as well as on a dataset of over 6 million of geotagged images (their knowledge base) from all over the world. Unknown images are described by se-

lected image descriptors (e.g., color histograms, GIST) and compared to the big knowledge base. The top-k most similar returned geotagged images are used to estimate the location of a given unknown image. Although this strategy is not precise in finding an exact location most of the time, it indicates roughly where an image was captured. For 16% of the time their method correctly predicted an image location to within 200km. Extensions of this approach rely solely on the text tags associated with the images [35]. Other work on photos' geotagging based solely on their visual content has emerged mostly for landmark recognition [23], but Kalantidis et al. [11] propose geotagging non-landmark images using a big geotagged and clustered dataset.

Strategies similar to the Hays and Efros' approach have been employed in the Placing Task, one of the tasks launched in 2010 at a benchmarking initiative to evaluate a "new algorithm for multimedia access and retrieval" (a spin-off of VideoCLEF), called MediaEval [17]. Some of its results (visual feature based) and its dataset will be present in Section 3 and Section 5.1.

3. PLACING TASK AT MEDIAEVAL

Placing Task requires participants to automatically assign latitude and longitude coordinates to each of the provided test videos. The most recent approaches for video geocoding were submitted to the Placing Task of MediaEval 2010 and 2011. They can be basically divided into methods based on textual information and methods based on visual information. Our interest in this paper is to compare with the methods based only on visual information, which were more frequent in the 2011 version of MediaEval Placing Task than in 2010.

In 2010 the Placing Task, there were three main approaches, as summarized in [17]: (a) geoparsing and geocoding texts extracted from metadata assisted by a gazetteer of geographic name, such as GeoName; (b) propagation of the georeference of a similar video in development database to the test video; (c) division of the training set in geographical regions determined by clustering or fixed-size grid using a model to assign items to each group. The model estimation was based on textual metadata and visual clues. The best result in 2010 for this task was accomplished by VanLaere et al. [41], using only metadata of images and videos, combining approaches (b) and (c): first a language model was employed to identify the most likely area of the video and then the most similar resources from the training set were used to determine the exact coordinates.

Just one research team reported results using only visual content in 2010 [13]. That work used visual features of the development set for training a multi-class Support Vector Machine (SVM) classifier with Radial Basis Function (RBF) kernel. Their best results were achieved by a hierarchical clustering with a diameter threshold of 100 km, which determined 317 classes for the SVM with the descriptors Color and Edge Directivity Descriptor (CEDD), Fuzzy Color and Texture Histogram (FCTH), and Gabor.

In 2011, four groups submitted for a run in which only visual features could be used to predict the location of the test videos. Most of them considered visual features as a backup predicting approach for the cases in which no tags or textual description associated with a test video.

Using an algorithm to compare video sequences [1], Li et al. [20] (UNICAMP team), concentrated only on visual features of a video to predict its location. None of the photos or key frames were used in this case. For each frame of an input video, motion features were extracted from the video stream. Videos are then compared by taking into account their motion features. Each video in the test set was compared with those in the development set. Then, for each test video, an ordered list of similar videos from the development set was produced along with its similarity score to that given test video. Finally, the most similar video of this list was picked as the one for transfering its known latitude/longitude to the query test video.

Choi et al. [7] (ICSI team) proposed an approach based on visual similarity between query video and items in development set, either videos keyframes or Flickr photos. However they extracted GIST features of frames and photos and ran 1-nearest neighbor to match each test video (its temporal mid-point frame) against the whole development set. The most similar/close (Euclidean distance) returned item was selected to give its latitude/longitude to the query video.

Hauff and Houben [8] (WISTUD team) divided the world globe in cells of variable size (small for dense data area and larger if sparse data) and assigned items from development set to their respective cells. For the visual approach, only 10% of the set was used. Matches between the query video and the videos of the training set work as follows: first the cell, C_{max}, with the highest probability to contain a test video is identified. Then an item inside C_{max} that is the closest match to test video is identified and its location is assigned to the test video. Those matches are implemented by a Naïve-Bayes nearest neighbor approach.

The strategy of Van Laere et al. [15] (UGENT team) was based on a language model and the Jaccard similarity search on textual tags associated with videos and photos. However, for visual similarity, they compared photos from the development set with keyframes of query videos, both represented by CEDD. If different keyframes of a video are most similar to different photos, a pair (keyframe, photo) with the highest degree of similarity is used to indicate the location of query video. Once the most similar photo (p) to the query video (v) is found, location of p is used as the prediction for the location of v.

4. BAG-OF-SCENES

In this section, we describe a novel model for video representation that is based on a dictionary of scenes[1]. In the scenario of video geocoding, the motivation for using this approach is that video frames are like pictures from places and these pictures have important information regarding the place location. If we have a dictionary of representative pictures from different places, we can describe video frames by considering their similarities to the representative pictures. Therefore, if a video has frames similar to photos taken in certain locations, we can infer that it is from such a location, facilitating the geocoding task. Given an input video, we create a vector of activations of each video frame to each of the scenes in the dictionary.

The most important advantage of the representation based on the dictionary of scenes is that it relies on semantic elements. Traditional dictionaries of local low-level descriptions, like SIFT or STIP, are composed by visual words

[1]In this paper, the term *scene* refers to images (photos), differently of its designation in video segmentation tasks.

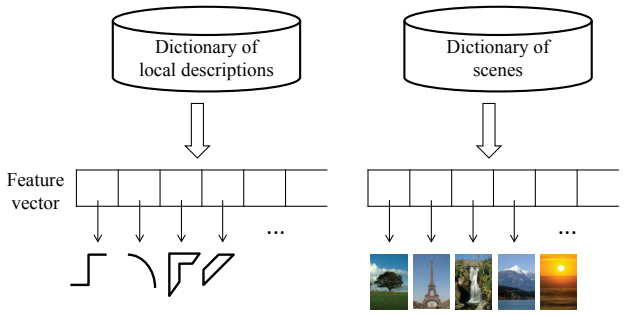


Figure 1: Comparison between the proposed dictionary of scenes to a dictionary based on local descriptions. We can notice that the representation based on the local dictionary relies on elements without clear semantics, like small corners and edges, while, the representation based on the dictionary of scenes carries much more semantics. In addition, the feature space for the dictionary of scenes has semantics in each dimension independently.

based on very punctual elements, like small corners and edges, which carry no semantic information [14]. The dictionary of scenes is composed by pictures and those pictures have more semantic information than corners and edges. Therefore, our final video representation is an activation vector to high-level elements, resulting in a representation space where each vector dimension has semantics itself. Figure 1 shows the differences between those types of dictionaries.

To generate a dictionary of scenes, we first need to compute a representation for each scene. Given a set of scenes which may come from frames of training set videos or from an arbitrary collection of images, each scene can be represented by a certain type of low-level feature, like color histograms or bag-of-visual-words, for example. Figure 2 illustrates the steps for generating a dictionary of scenes and the steps to represent a video using the dictionary. The visual dictionary is created by selecting feature vectors of the scenes according to some criteria. One can cluster the feature space in the same fashion it is performed for SIFT dictionaries [36, 39, 40, 42]. Other possibilities rely on a random selection of scenes or even on a manual selection of the most important scenes for the target application. In our application scenario, a selection of representative scenes from places of interest may be more promising.

It is important to highlight that any technique can be used for frame extraction from videos, like sampling at fixed-time intervals or by employing summarization methods [2–4].

Another important aspect of the description based on dictionaries, and also valid for the dictionary of scenes, is that the feature vectors of each scene and the feature vectors of each visual word need to be of the same nature. In our case, a visual word is also a scene. For example, if we generate the dictionary by representing the scenes with a 64-bin color histogram, each video frame considered in the dictionary also need to have a 64-bin color histogram representation.

Once the dictionary is generated, we are able to create a high-level representation for videos. Assignment approaches are then used to describe the feature vector of each frame according to the dictionary. The *hard* and *soft* assignment methods, popularly used with SIFT dictionaries [21, 31, 40]

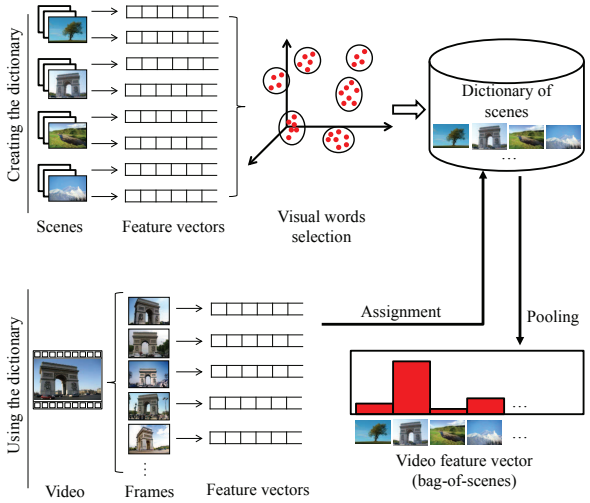


Figure 2: The schema for generating and using a dictionary of scenes. The dictionary is created based on a given collection of scenes, which may come from an image dataset or from video frames. After representing each image with any kind of feature vector, some of them are selected to compose the dictionary. Given an input video to be represented, its frames are assigned to one or more of the scenes in the dictionary. A pooling strategy is then applied to generate the video feature vector (*bag-of-scenes*).

are suitable in this step. To generate the final *bag-of-scenes* representation for a video, we can employ pooling strategies, like the popular *average* and *max* pooling [6]. The second part of Figure 2 shows the steps for generating the representation based on the dictionary of scenes.

The bag-of-scenes representation has some interesting properties. As the visual words are scenes, which tend to carry semantic information, the activation vector has one position for each concept, making it simple to analyze the presence or absence of each concept into a video. In the video geocoding scenario, the feature vector is a *place* activation vector, because each visual word is a picture of some specific place. Mathematically speaking, the dictionary of scenes creates a vector space where each dimension represents a specific semantic concept. It is important to realize that, despite our dictionary of scenes is being originally proposed and validated for video geocoding, it can be applied to many other applications, like video categorization or video retrieval, for instance.

5. EXPERIMENTS

The goal of the experiments is to evaluate the dictionary of scenes model for video geocoding. To create a suitable scenario, we have worked under all the specifications of the placing task of MediaEval 2011. The details of the task as well as the datasets used, are explained in Section 5.1. Our strategies to create and employ the proposed model to solve the task are presented in Section 5.2.

5.1 Datasets & evaluation criteria

Participants in the Placing Task at MedialEval 2011 were allowed to use image/video metadata, audio and visual fea-

tures, as well as external resources, depending on the run submitted. The organizer of this task released two sets of data [33]. The first set is meant to the development and training of algorithms, thus called development set. It is comprised of geotagged Flickr videos as well as the metadata for geotagged Flickr images, such as title, tags, and descriptions provided by the owner of that resource, comments of her/his friends, users' contact lists, and other uploaded resources on Flickr. This development data included 10,216 geotagged videos, along with their extracted keyframes and corresponding pre-extracted low-level visual features, and metadata. For only development and training purposes, this set also included visual features and metadata for 3,185,258 CC-licensed Flickr photos, uniformly sampled from all parts of the world. The latitude and longitude of those videos and photos were also informed.

The second set, called test data, is composed by solely 5,347 videos, their keyframes with extracted visual features and related metadata (without geographic location).

Keyframes were extracted at each 4 second intervals from videos and saved as individual JPEG-format images. The following visual feature descriptors for keyframes and photos were extracted and provided: Color and Edge Directivity Descriptor (CEDD), Gabor Texture, Fuzzy Color and Texture Histogram (FCTH), Color Histogram, Scalable Color, Auto Color Correlogram, Tamura Texture, Edge Histogram, and Color Layout.

Participants in placing task were required to submit at least one run that uses only audio/visual features. The result evaluation was based on the distance to the ground truth geographic coordinate point, in a series of widening circles of radius (in km): 1, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000. Thus, an estimated location is counted as correct at a particular circle size, which can be seen as quality or precision level, if it lies within a given circle radius.

More details about Placing Task at MediaEval 2011 are given at the working notes of the task's organizer [33].

5.2 Experimental setup

Our experiments are divided into two phases. The first phase comprises of the parameter adjustments using the development set. The second phase employs the best dictionary configurations for representing and geocoding videos from the test set. In each of the phases, we have used two sources for the scenes to generate the dictionary: videos frames from the development set and Flickr photos. To easily distinguish between them, in the remainder of this section, we call the former as *dictionary of frames* and the latter as *dictionary of scenes*.

As explained in Section 5.1, each video from the MediaEval 2011 dataset is accompanied by a set of keyframes. In addition, each keyframe also has a set of global low-level feature vectors computed. The more than 3 million geotagged Flickr images supplied by MediaEval 2011 also have a set of global low-level features already extracted. In our experiments, we have used many of the low-level descriptions provided, trying to discover which of them are better for the placing task.

All of our dictionaries were generated at random, which means that we have randomly selected frames from the development set or scenes from the Flickr dataset to be the visual words of the dictionary. For SIFT-based dictionaries, a random selection of visual words has similar performance to clustering techniques, due to the curse of the dimensionality [42]. In the dictionary of scenes, the dimensionality is still a problem. If we can obtain good results even with this simple way of generating the dictionary of scenes, it is an indication of how promising the idea is.

To represent videos by a given dictionary of scenes, we have employed the state-of-the-art assignment and pooling techniques of the image representation community [6,39,40]. Hard and soft assignment as well as average and max pooling were used. Details of these techniques are presented in Section 2.2.

After computing the bag-of-scenes representation for each video, our strategy to assign a global location for a given video is based only on the visual information. We have computed the Euclidean distance from a query video to all the remaining videos in the development and estimated its latitude/longitude by assigning those from the nearest video location. The evaluation measures were computed using the distance circles to the correct coordinate point, as explained in Section 5.1. Our results were not submitted to the Placing Task at MediaEval 2011, however, official evaluation was possible by running the official evaluation program, which was released for participant groups after the event.

5.3 Results on the development set

The experiments in the development set combine different parameters for creating and using the dictionary. To evaluate the parameters, we have used all the videos in the development set as queries and, when estimating their latitude/longitude by assigning the location of the nearest video, we considered that the query video was not part of the development set. Our analysis using the *dictionary of frames* has shown that a good configuration for the visual dictionary uses CEDD descriptor, soft assignment ($\sigma = 3$), and max pooling. The soft assignment and max pooling implementations follow the equations provided in Section 2.2. Although other σ values were also tested, $\sigma = 3$ was selected because it makes a frame to be assigned to a fair number of visual words, considering the CEDD feature space. There was few impact when changing the dictionary size. A meaningful difference occurred when using a very small or a very large dictionary, 30 and 50,000 visual words, but they were worse than dictionaries of sizes 50, 500, and 5000. The experiments with the *dictionary of scenes* in the development set also shows that CEDD descriptor, soft assignment ($\sigma = 3$), and max pooling achieve the best results. We have tried dictionaries up to 50,000 visual words, but the results were better with smaller dictionaries.

Table 1 presents those results and compares the two types of dictionary. We can note that there is a few difference between the *dictionary of frames* and the *dictionary of scenes*.

Table 1: Experiment results showing small performance difference between dictionary of *frames* and dictionary of *scenes* in the development set. The values are the percentage of videos from the development set that were correctly geocoded in the radii 1km, 10km, and 100km.

Dictionary	% 1km	% 10km	% 100km
Frames	14.59	15,69	17.23
Scenes	13.60	14,62	16.15

This is an interesting result, because frames are clearly elements that came from the same dataset, while the scenes came from a completely different source. It opens up a number of possibilities that deserve much deeper study, but an immediate consequence is that we can create a good dictionary even with a kind of information that comes from a completely unrelated source. This phenomenon has been a trend in the machine learning community, known as *transfer learning* [29].

5.4 Results on the test set

According to the experimental results on the development set, we have used CEDD descriptor, soft assignment ($\sigma = 3$), and max pooling to run the experiments on the test set. The implementations of soft assignment and max pooling are the same of the previous experiments and follow the equations in Section 2.2. We have tested 3 different dictionary sizes: 50, 500, and 5000. The dictionaries were created using frames from the development set, in the case of the *dictionary of frames*, and using Flickr images for the *dictionary of scenes*.

The results for the *dictionary of frames* and the *dictionary of scenes* in the test set are shown in Figure 3(a) and (b), respectively. We can note that, the variation in dictionary sizes has few impact in the results. One possible reason for the dictionaries sizes not affecting the results considerably is that the random selection of visual words may have taken many images with few information about place location. Hence, the small portion of representative visual words helped the geocoding of only some of the test videos.

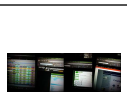
Another result observed in the test set is the small difference between using a *dictionary of frames* or a *dictionary of scenes*, as also observed in the experiments in the development set. One reason may be the large number of non-informative visual words, that occurred in the random selection of both scenes and frames.

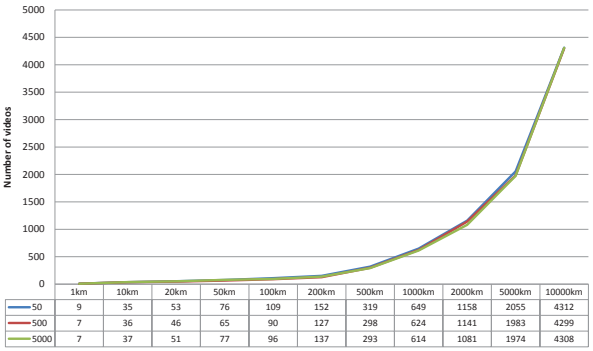
To evaluate the quality of the representation when using the *dictionary of scenes*, we have verified the visual words activated by the videos that we tagged correctly. The most activated scenes by the best geotagged videos are shown in Table 2. Notice that, despite those videos were tagged really close to the correct location, the scenes activated by them are not necessarily representative from the location. It is important to note that, the scenes themselves do not need to be specifically from a location. However, videos that are specifically from a certain location should activate the same scenes. What might have happened in the case of the best tagged test videos is that, there are videos in the development set which are from the same location and have activated the same scenes from the dictionary.

Table 3 compares the results obtained by the proposed method with those reported by four participants of the MediaEval 2011 Placing Task: UGENT [15], UNICAMP [20], ICSI [7], and WISTUD [8]. They were the ones to consider methods based only on the visual information. We can see that our approach performs better than most of the compared methods, except for that of the UNICAMP team [20]. This method is based on motion information and, hence, it does not consider visual properties of video frames in an independent manner. Such a method has geotagged correctly videos that our approach tagged wrongly and vice versa.

Although the proposed method is not superior to the state-of-the-art approaches for video geocoding, the obtained results show the potential of the idea. Observe that, by generating a video representation based only on pictures, which come from a completely different source in the case of the *dictionary of scenes*, it is still enough to provide a good representation for video geocoding. Despite our very simple way to generate the visual dictionary, which has taken pho-

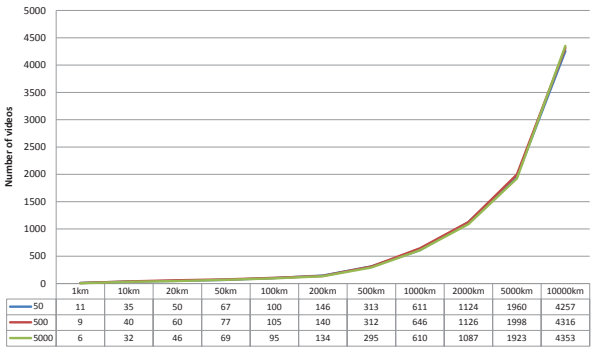
Table 2: Ten most activated visual words by some of the best geotagged videos when using the dictionary of 5000 scenes. The value below the video thumbnail is its distance to the correct location, while the value below each visual word is its activation value, in percentage, by the corresponding video.

Video	1	2	3	4	5	6	7	8	9	10
										
0.004	12.5	6.4	6.4	6.0	5.8	5.8	5.7	4.3	4.3	3.6
									
0.012	1.7	1.7	1.6	1.3	1.3	1.2	1.1	1.1	1.0	1.0
										
0.516	4.1	3.4	2.5	1.8	1.7	1.7	1.6	1.5	1.5	1.5
										
0.603	0.9	0.9	0.8	0.8	0.8	0.8	0.8	0.8	0.7	0.7
0.861	1.3	1.2	1.1	1.0	1.0	1.0	1.0	1.0	0.9	0.8



	1km	10km	20km	50km	100km	200km	500km	1000km	2000km	5000km	10000km
50	9	35	53	76	109	152	319	649	1158	2055	4312
500	7	36	46	65	90	127	298	624	1141	1983	4299
5000	7	37	51	77	96	137	293	614	1081	1974	4308

(a)



	1km	10km	20km	50km	100km	200km	500km	1000km	2000km	5000km	10000km
50	11	35	50	67	100	146	313	611	1124	1960	4257
500	9	40	60	77	105	140	312	646	1126	1998	4316
5000	6	32	46	69	95	134	295	610	1087	1923	4353

(b)

Figure 3: Results using (a) the *dictionary of frames* and (b) the *dictionary of scenes* in the test set. The values are the number of test videos correctly geocoded at different distances from the correct video location. We can see that the change in dictionary size causes almost no difference in the quality of the representation.

Table 3: Comparison of the results obtained by the proposed approach with those reported by four participants of the MediaEval 2011 Placing Task. The values are the number of test videos correctly geocoded at different distances from the correct video location.

Radius					Dictionary of Frames			Dictionary of Scenes		
(km)	UGENT [15]	UNICAMP [20]	ICSI [7]	WISTUD [8]	50	500	5000	50	500	5000
1	2	11	5	0	9	7	7	11	9	6
10	6	60	16	5	35	36	37	35	40	32
100	49	145	67	-	109	90	96	100	105	95
1000	624	650	598	583	649	624	614	611	646	610
10000	4332	4248	4234	-	4312	4299	4308	4257	4316	4353

tos at random, the results are comparable to (or even better than) some of the methods presented in Table 3.

We can think about ways of improving the bag-of-scenes model. Our random selection of pictures to compose the dictionary may take pictures with very few information regarding the place location and, thus, being no informative for the placing task. Notice that some of those non-informative pictures were activated even in our best geotagged videos, as shown in Table 2. A smarter selection of scenes may be able to create more informative dictionaries and, hence, improve the video representation for geocoding.

6. CONCLUSIONS

This paper has introduced a new video representation for visual-based video geocoding, named bag-of-scenes. This representation model relies on a dictionary of scenes, whose visual words carry more semantic information than local low-level features, like SIFT. The feature space spanned by such a model has the property of having one dimension for each semantic concept. By generating a dictionary of representative scenes from places of interest, we can create a high-level representation for video geocoding.

Our experiments were conducted in the context of the Placing Task at MediaEval 2011. Despite our simple strategy for creating the dictionary of scenes, based on a random selection of pictures, the results have shown that our approach performs similar to most of the methods submitted to the Placing Task at MediaEval 2011.

Future work includes the investigation of smarter procedures for selecting informative scenes to be used in the creation of visual dictionaries. We also plan to investigate new approaches that exploit transfer learning in geocoding tasks.

7. ACKNOWLEDGEMENTS

We would like to thank Eduardo Valle for their valuable contributions. This research was partially supported by AMD, Microsoft Research and Brazilian agencies FAPESP (grants 2007/52015-0, 2009/10554-8, 2009/05951-8, 2009/18438-7 and 2011/11171-5), CNPq (grant 306587/2009-2), and CAPES.

8. REFERENCES

[1] J. Almeida, N. J. Leite, and R. Torres. Comparison of video sequences with histograms of motion patterns. In *ICIP*, pages 3673–3676, 2011.

[2] J. Almeida, N. J. Leite, and R. Torres. VISON: VIdeo Summarization for ONline applications. *Pattern Recognition Letters*, 33(4):397–409, 2012.

[3] J. Almeida, N. J. Leite, and R. Torres. Online video summarization on compressed domain. *J. Visual Communication and Image Representation*, 2012. DOI: 10.1016/j.jvcir.2012.01.009.

[4] J. Almeida, R. Torres, and N. J. Leite. Rapid video summarization on compressed video. In *ISM*, pages 113–120, 2010.

[5] S. Avila, N. Thome, M. Cord, E. Valle, and A. de A. Araújo. Bossa: Extended bow formalism for image classification. In *ICIP*, pages 2966–2969, 2011.

[6] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. *CVPR*, pages 2559–2566, 2010.

[7] J. Choi, H. Lei, and G. Friedland. The 2011 ICSI video location estimation system. In *Working Notes Proc. MediaEval Workshop*, volume 807, 2011.

[8] C. Hauff and G.-J. Houben. WISTUD at MediaEval

2011: Placing task. In *Working Notes Proc. MediaEval Workshop*, volume 807, 2011.

[9] J. Hays and A. A. Efros. im2gps: estimating geographic information from a single image. In *CVPR*, 2008.

[10] Y.-G. Jiang and C.-W. Ngo. Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval. *Computer Vision and Image Understanding*, 113(3):405–414, 2009.

[11] Y. Kalantidis, G. Tolias, Y. Avrithis, M. Phinikettos, E. Spyrou, P. Mylonas, and S. Kollias. Viral: Visual image retrieval and localization. *Multimedia Tools and Applications*, 51:555–592, 2011.

[12] Y. Ke, R. Sukthankar, and L. Huston. An efficient parts-based near-duplicate and sub-image retrieval system. In *ACM MM*, pages 869–876, 2004.

[13] P. Kelm, S. Schmiedeke, and T. Sikora. Multi-modal, Multi-resource Methods for Placing Flickr Videos on the Map. In *ACM ICMR*, 2011.

[14] E. P. X. L-J. Li, H. Su and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.

[15] O. V. Laere, S. Schockaert, and B. Dhoedt. Ghent university at the 2011 placing task. In *Working Notes Proc. MediaEval Workshop*, volume 807, 2011.

[16] I. Laptev. On space-time interest points. *Int. J. Comp. Vision*, 64(2–3):107–123, 2005.

[17] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and G. J. F. Jones. Automatic tagging and geotagging in video collections and communities. In *ACM ICMR*, pages 51:1–51:8, 2011.

[18] R. R. Larson. Geographic information retrieval and digital libraries. In *ECDL*, volume 5714/2009, pages 461–464, 2009.

[19] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2, pages 2169–2178, 2006.

[20] L. T. Li, J. Almeida, and R. Torres. RECOD working notes for placing task MediaEval 2011. In *Working Notes Proc. MediaEval Workshop*, volume 807, 2011.

[21] L. Liu, L. Wang, and X. Liu. In defense of soft-assignment coding. In *ICCV*, pages 1–8, 2011.

[22] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comp. Vision*, 60(2):91–110, 2004.

[23] J. Luo, D. Joshi, J. Yu, and A. Gallagher. Geotagging in multimedia and computer vision–a survey. *Multimedia Tools Appl.*, 51:187–211, 2011.

[24] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Trans. Circuits Syst. Video Techn.*, 11(6):703–715, 2001.

[25] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. J. Comp. Vision*, 60(1):63–86, 2004.

[26] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *TPAMI*, 27(10):1615–1630, 2005.

[27] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. J. Van Gool. A comparison of affine region detectors. *Int. J. Comp. Vision*, 65(1-2):43–72, 2005.

[28] C.-W. Ngo, W. Zhao, and Y.-G. Jiang. Fast tracking of near-duplicate keyframes in broadcast domain with transitivity propagation. In *ACM MM*, pages 845–854, 2006.

[29] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE TKDE*, 22(10):1345–1359, 2010.

[30] O. A. B. Penatti, E. Valle, and R. Torres. Encoding spatial arrangement of visual words. In *CIARP*, volume 7042, pages 240–247, 2011.

[31] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, pages 1–8, Jun. 2008.

[32] M. J. Pickering, D. Heesch, S. M. Rüger, R. O'Callaghan, and D. R. Bull. Video retrieval using global features in keyframes. In *TREC*, 2002.

[33] A. Rae, V. Murdock, P. Serdyukov, and P. Kelm. Working notes for the placing task at MediaEval 2011. In *Working Notes Proc. MediaEval Workshop*, volume 807, 2011.

[34] M. Rautiainen and D. S. Doermann. Temporal color correlograms for video retrieval. In *ICPR*, pages 267–270, 2002.

[35] P. Serdyukov, V. Murdock, and R. van Zwol. Placing flickr photos on a map. In *ACM SIGIR*, pages 484–491, 2009.

[36] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477 vol.2, 2003.

[37] J. R. Smith, S. Srinivasan, A. Amir, S. Basu, G. Iyengar, C.-Y. Lin, M. R. Naphade, D. B. Ponceleon, and B. L. Tseng. Integrating features, models, and semantics for trec video retrieval. In *TREC*, 2001.

[38] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and Trends in Computer Graphics and Vision*, 3:177–280, 2008.

[39] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *TPAMI*, 32(9):1582–1596, 2010.

[40] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *TPAMI*, 32:1271–1283, 2010.

[41] O. Van Laere, S. Schockaert, and B. Dhoedt. Finding locations of flickr resources using language models and similarity search. In *ACM ICMR*, pages 48:1–48:8, 2011.

[42] V. Viitaniemi and J. Laaksonen. Experiments on selection of codebooks for local image feature histograms. In *Int. Conf. on Visual Inf. Systems: Web-Based Visual Inf. Search and Management*, pages 126–137, 2008.

[43] L. Wu, Y. Guo, X. Qiu, Z. Feng, J. Rong, W. Jin, D. Zhou, R. Wang, and M. Jin. Fudan university at TRECVID 2003. In *TRECVid*, 2003.

[44] X. Wu, W. Zhao, and C.-W. Ngo. Near-duplicate keyframe retrieval with visual keywords and semantic context. In *CIVR*, pages 162–169, 2007.