Automatic Tagging and Geotagging in Video Collections and Communities *

Martha Larson
Delft University of Technology
Delft, the Netherlands
m.a.larson@tudelft.nl

Mohammad Soleymani
University of Geneva
Geneva, Switzerland
mohammad.soleymani@unige.ch

Pavel Serdyukov
Yandex
Moscow, Russia
pavser@yandex-team.ru

ABSTRACT

Automatically generated tags and geotags hold great promise to improve access to video collections and online communities. We overview three tasks offered in the MediaEval 2010 benchmarking initiative, for each, describing its use scenario, definition and the data set released. For each task, a reference algorithm is presented that was used within MediaEval 2010 and comments are included on lessons learned. The *Tagging Task, Professional* involves automatically matching episodes in a collection of Dutch television with subject labels drawn from the keyword thesaurus used by the archive staff. The *Tagging Task, Wild Wild Web* involves automatically predicting the tags that are assigned by users to their online videos. Finally, the *Placing Task* requires automatically assigning geo-coordinates to videos. The specification of each task admits the use of the full range of available information including user-generated metadata, speech recognition transcripts, audio, and visual features.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*

General Terms

Algorithms, Measurement, Performance

1. INTRODUCTION

A familiar approach to video annotation involves associating a video content item (a clip, a shot or an episode of a program) with terms that reflect its content and subject matter. This basic act of labeling is used in large multimedia archives, where professional archivists pick subject category names from a controlled vocabulary in order to annotate video. It is also used on the Internet, where the public at large chooses characteristic words or phrases as 'tags' to annotate video. These annotation forms are similar in that they support users in finding and re-finding video content. Increasingly, users also specify the location of video in the form of geo-coordinates, also known as geotagging. Despite the widespread adoption of tagging practices, much content within private collections and also within multimedia communities on the Internet remains untagged. Automatic techniques capable of generating tags and subject labels have an enormous potential to improve the performance of multimedia retrieval, complementing user-contributed tags and providing users with improved access to multimedia content.

In this paper, we present three tasks devoted to tagging and geotagging offered by the MediaEval 2010 benchmarking initiative. MediaEval (http://www.multimediaeval.org) brings multimedia researchers together to pool research resources and focus effort on developing solutions for challenging issues facing multimedia indexing and retrieval. The specific goal is to support research in the area of multimedia access and retrieval that is focused on speech, language and contextual (especially geographical and social) aspects of video. In order to promote reproducibility and re-usability, the MediaEval initiative licenses the data sets for public use. For each task, we give the task definition, describe the data set and present a reference algorithm developed by participants in the MediaEval 2010 benchmark, who are represented among the authors of this paper. The reference algorithm provides an indication of the performance level to be beat in order to achieve overall improvement on the task.

This paper is organized as follows. After presenting related work in the next section, we cover each of the three tasks in turn: the *Tagging Task, Professional*, the *Tagging Task, Wild Wild Web*, and the *Placing Task*. We finish with concluding comments and an outlook onto MediaEval 2011.

2. RELATED WORK

This section provides a brief overview of relevant literature on automatic tagging, including supporting information on related work in speech indexing, geotagging and benchmarking initiatives related to multimedia retrieval and indexing.

2.1 Tagging and Subject Label Generation

The purpose of assigning labels to content is to achieve a representation that encodes a higher level of semantic abstraction. Within multimedia information retrieval systems,

*This paper was written collaboratively by the organizers of the MediaEval 2010 benchmarking initiative. Please refer to the final section of the paper for a list of the names of the other authors. The third author was affiliated with Delft University of Technology during MediaEval 2010.

ICMR '11, April 17-20, Trento, Italy

indexing features provide high-level semantic generalizations concerning individual items. Automatic approaches to label or tag prediction generally follow one of the two approaches, distinguished by [11]. The first approach involves *extraction*. Here, appropriate labels for items are chosen from among the words or phrases already associated with item content or metadata. Under this approach, the identity of the labels to be assigned need not be known in advance, a task we refer to as 'open-set tagging'. The second approach involves *assignment*. Here, labels from a fixed set of labels are assigned to items, a task we refer to as 'closed-set tagging'.

The task of closed-set tagging can be approached either as a series of binary classification tasks, one for each label, or as a multi-class classification task. Labeling is thus effectively an item categorization problem, such as addressed by [9]. Recently, an information retrieval (IR) approach has been adopted for the task of closed-set tagging. Here, the tag or subject label is treated as a query and used to query a collection consisting of the documents to be annotated, e.g., [25]. The advantage of this approach is that no training data is needed. Conventional IR query expansion methods can be used to expand the class labels into appropriately enriched queries. When this approach is operationalized, a cutoff point—which is possibly label dependent—is defined in the list of items returned by the system. The tag used as a query is then assigned to all items above the cutoff.

Increasingly, social and other contextual information is being exploited for tag prediction. In [16], tags are predicted for bookmarked URLs using page text, anchor text, linked websites, and tags of other URLs. Different sources of information have successfully been integrated in factorization models to predict the tags that a user will assign to an item [30].

Parallel to the development of modern speech recognition systems, researchers have addressed the task of automatically generating labels that characterize the content of spoken documents. The first work, dating from the early nineties, addressed the task of classifying speech messages into one of six topic classes [31].

Automatic spoken audio labeling often attempts to reproduce labels that would be generated by a human, given a specific classification scheme. For example, in [27], an algorithm is presented that uses a classifier to predict class-labels for television news that are drawn from the Media Topic taxonomy of the International Press Telecommunications Council (http://www.iptc.org). Example classes are: *politics*, *sport* and *lifestyle and leisure*. The classifier exploits word distributions characteristic of particular topics to assign items to classes. The decision of the classifier is made difficult by the presence of noise in the transcripts in the form of speech recognition errors. Note that a spoken content item about a given topic, will probably not mention the subject label of that topic explicitly. For example, a review of a new restaurant would fall into the category of *lifestyle and leisure*, but would be unlikely to contain either of the words 'lifestyle' or 'leisure'. This example illustrates why a keyword extraction approach is not particularly well suited for generating labels that represent content at a high level of abstraction: It is not given, or even expected, that words that reflect membership in a high-level category are used in the spoken content of an item. In [26], a classification approach is proposed that automatically assigns terms to spoken interview recordings drawn from a thesaurus developed by scholars specialized in the domain. Uses for terms include representing items in the interface and expanding queries. The Tagging Task, Professional, subject labels have similar applicability.

Research effort has also been devoted to prediction of genre labels for multimedia. For example, [24] proposes a multimodal approach that uses speech transcripts to supplement visual features. Category schemes used for genre are domain dependent and can be expected to vary from one use scenario to the next. In general, however, a genre category combines elements of style, form and topic. Conventionally, topic plays a major role in genre categories, as exemplified by the choice of genre categories used in [24]: *football*, *cartoons*, *music*, *weather forecast*, *newscast*, *talk show* and *commercials*.

Not all tags important for multimedia tagging are related to topic, however. Much work has been devoted to automatic classification approaches that tag aspects of content items that are more closely related to function, form or style. In [35], an approach is created for predicting categories of spoken content units occurring in human telephone conversations. The categories include *statement*, *question* and *apology*. In [8], a set of concepts to be used for search and navigation is extracted from call center recordings using models built of linguistic rules. The set of labels assigned by the system includes categories such as *positive contact* and *negative contact* reflecting whether or not the caller was satisfied with the call center interaction. In Section 4.2, it will be seen that the tag set of the Tagging Task, Wild Wild Web, includes a mix of topical and non-topical tags.

2.2 Automatic Geotagging of Multimedia

Previous work that has been carried out in the area of automatic geotagging of multimedia has focused on images, usually from Flickr. User-contributed tags have a strong location component, as brought out by [33], who reported that over 13% of Flickr image tags could be classified as locations using Wordnet. In [29], the geo-locations associated with specific Flickr tags are predicted using spatial distributions of tag use. A tag which is strongly concentrated in a specific location has a semantic relationship with that location. Understanding that a tag (or a term) is highly correlated with a place is the key to understand whether a user has a geographic intent in mind when tagging an image or searching for images on the Web. User-contributed tags are exploited for geotagging by [32], who use tag distributions associated with locations represented as grid cells on a map of the Earth is used to infer the geographic locations of Flickr images.

Tag prediction methods exploiting the visually depicted content of images include [14], which uses visual features and a nearest-neighbor classification method to geotag Flickr images. The data set is, however, limited to a subset of Flickr images tagged with at least one name of a country, continent, densely populated city or popular tourist site and not tagged with specific non-geographic tags such as `birthday` or `concert`. Visual, textual and temporal features are combined by [7], which investigates the classification of images within specific cities. For 100 cities, the top ten landmarks in that city are identified and 10-way classification of photos geotagged around these landmarks is performed. Other closely related work addresses geotagging of non-multimedia objects, for instance finding the geographical focus of Web pages [2] or short Twitter messages [4].

2.3 Other Benchmarking Initiatives

Benchmarking and benchmark data sets serve to concentrate research effort, enable cross-site comparison and drive forward the state of the art. The stronghold of information retrieval benchmarking is TREC, the Text REtrieval Conference (http://trec.nist.gov) established in 1992 by the US National Institute of Standards and Technology (NIST). The first major spoken-content-based benchmark, TREC Spoken Document Retrieval (TREC-SDR) [12] was devoted broadcast news retrieval and ran from 1997-2000. TREC-SDR made use of data from the NIST Topic Detection and Tracking [1] campaign, which ran tasks on a broadcast news corpus from 1998-2004. The retrieval of spoken content was then picked up by CLEF, the Cross Language Evaluation Forum in Europe (http://www.clef-campaign.org). CLEF-SDR (2003-2004) was followed by CLEF-SR (2005-2007), a Speech Retrieval track [28] that moved evaluation beyond broadcast news content to more challenging collections consisting of a speech stream rather than well structured spoken documents. CLEF-SR was followed by VideoCLEF, which ran several tasks, including subject label prediction, in 2008-2009. In 2010, VideoCLEF expanded its task offering and became an independent benchmarking initiative with the name MediaEval.

A close 'big cousin' of MediaEval is TRECVid [34], a video retrieval benchmark that ran as a TREC track 2001-2002 and in 2003 became an independent benchmark. Earlier on, TRECVid devoted considerable attention to broadcast news content and then turned to other content, including television programming from NISV, the Netherlands Institute for Sound and Vision (http://instituut.beeldengeluid.nl). Traditionally, TRECVid has focused on what is depicted in the visual channel of the video at the shot level. In contrast, MediaEval treats video units of varying sizes and is interested in the overall meaning of the video, including its topical content and context.

Recently, video classification has attracted renewed interest in a related form, namely the genre classification task set out by Google as an ACM Multimedia Grand Challenge task in 2009 and 2010. Finally, the ECML PKDD Discovery Challenge (2008-09) ran a tag recommendation task involving Bibsonomy [17, 10]. Participants were supplied with bookmarks and bibtex files and required to predict tags.

3. TAGGING TASK PROFESSIONAL

The *Tagging Task, Professional* emulates the activity of a human archivist assigning subject labels to television broadcast content. Task participants are required to automatically match episodes in a collection of Dutch-language television broadcasts from the NISV archive with subject labels drawn from the keyword thesaurus used by the archive staff for annotation. This task is generally approached as 'closed-set tagging' meaning that the systems are given the identity of the subject labels in advance. It is a multi-label problem, meaning that a single video can have more than one label.

3.1 Use Scenario

With the growth of digital content flowing in at large multimedia archives such as NISV, where thousands of hours of content are archived every year, automatic analysis of multimedia content is a prerequisite for exploitation. Deploying (semi-) automatic annotation strategies could speed up the annotation work considerably and also allow for the annotation of content that would otherwise be left unannotated. At NISV, archivists use subject labels, keywords drawn from a conventionalized, but open vocabulary (Common Thesaurus Audiovisual Archives) to archive and retrieve videos. The use scenario for this task is semi-automatic tag recommendation. The subject classification task is derived from the archivist subject labeling use scenario.

3.2 Data Set and Evaluation

The data set contains television content in the Dutch language and is a mixture of various types of content, including news magazines, science news and documentaries. Thematic subject labels that have been assigned to the videos by the archive staff are used as ground truth. Provided with the data set are speech recognition transcripts [18] and shot-level information (shot boundaries plus one extracted keyframe per shot) [19]. The development set contains a large subset of the TRECVid 2007 and 2008 data sets. Videos lacking a critical element (e.g., they have no subject label) have been removed. The final development set consists of 405 videos and a set of 37 subject labels. These labels were selected such that each of them has more than 5 videos associated to them. The list of labels was post-processed by a normalization process that included standardization of the form of the label and elimination of labels encoding the names of personages or sources. The test set is composed of videos from TRECVid 2009 data set using the same selection criteria. The final test set contains 378 videos and 41 subject labels. The test set is mutually exclusive with the development set. Note that because we re-use the TRECVid data set, a large number of additional resources created by TRECVid (e.g., machine translations of transcripts and visual concept detection output) are available for use in this task.

3.3 Algorithm and Results

The algorithm, developed by Novay (http://www.novay.nl), addresses the task as a closed-set tagging problem. It applies an IR approach making use of the divergence model [21, 40]. The subject labels are treated as queries and the set of videos is treated as a document collection. The system assigns the label to those documents that are returned at top ranks. For evaluation purposes, we do not determine the cutoff that would be used in an operational setting. Rather, we use Mean Average Precision (MAP) to report the quality of the entire returned list of items. Although the model could be applied to any source of textual features, here, we demonstrate the basic principle using features derived only from the video metadata. As a baseline for the relevance $r(d, q)$ of an item d for a query term q we set $r(d, q) = p(q|d)$, where $p(q|d)$ is the number of occurrences of q in d divided by the total number of terms in d. The results of this baseline are given in the first line of the first column of Table 1. Only one out of the 41 labels (*kunstenaars*) does not occur as a word in any item and cannot be assigned under the baseline method.

The algorithm makes use of an item model (i.e., a document model) and a label model (i.e., a query model). Items are ranked in increasing order of their divergence from the query. Following [21], we use Markov chains to obtain these models. Given a document collection D, the language model for a query term q is defined as

Table 1: Performance on the *Tagging Task, Professional* (Mean Average Precision)

Method	no syn.	synonyms
Frequency	0.37	0.42
Divergence	0.42	0.47
Max. Entropy	0.43	0.48
Divergence (incl. dev. set)	0.45	0.48
Max. Entropy (incl. dev. set)	0.46	0.49

$$\bar{p}_q(t) = \sum_{d \in D} p(t|d)p(d|q), \quad (1)$$

where $p(t|d)$, the *term distribution* of d, is the probability that a term from d is an instance of t, and where $p(d|t)$, the *source distribution* of t, is the probability that a randomly selected occurrence of t has source d. The terms that are used are words extracted from the video metadata and the labels that belong to the open-class parts of speech (i.e., nouns, verbs, adjectives and adverbs) and have been mapped to their base forms (i.e., lemmata). Most labels used as queries in this task are given in plural form and are reduced to singular form for matching with the base forms. More details on the model can be found in [38, 39]. Assuming that q is a term like other terms, we also call this distribution the co-occurrence distribution of q.

The language model of a document d could simply be the distribution $p(t|d)$ of terms in document d. However, as usual we take a smoothed version, obtained by the same Markov chain and formally defined by

$$\bar{p}_d(t) = \sum_{d',t'} p(t|d')p(d'|t')p(t'|d). \quad (2)$$

For the comparison of the co-occurrence distribution and the document distribution we use the Jensen-Shannon divergence [6]. Results of this relevance measure are given in the first results column, second line of Table 1.

The evaluation for the task involves a ranking of documents for a given query. However, the annotators that assigned the original keywords selected terms relevant for a given document, not documents relevant for a given keyword. Since there is a potential mismatch between these two approaches, we include both the divergence and the rank of a keyword for given document in a linear model. The relevance of a document d for a query term q now becomes:

$$r(d|q) = \alpha + \beta\, p(q|d) - \gamma\, \mathrm{JSD}(\bar{p}_q, \bar{p}_d) - \delta\, \mathrm{rank}(q,d)/n_l, \quad (3)$$

where $\mathrm{JSD}(\overline{p_q}, \overline{p_d})$ is the Jensen-Shannon divergence between $\overline{p_w}$ and $\overline{p_d}$, $\mathrm{rank}(q,d)$ is the rank of q for d and n_l the total number of labels used. The coefficients were determined using a maximum entropy model on the test set ($\alpha = 1.0$, $\beta = 2.0$, $\gamma = 1.0$, $\delta = 0.17$). The results of this relevance measure are given in the first results column, third line of Table 1. As expected, this gives indeed a slight improvement over the run using only the divergence.

The co-occurrence distribution of a term can be seen as a proxy for its semantics. In this sense, the distribution will improve if we take more documents into account for the computation of the co-occurrence probabilities. Thus, in the next two runs we have used the abstracts from the test and the development set to compute the co-occurrence distributions. Again, we can use only the divergence or combine it with the other features. The results of these two runs are given in the last two lines, again showing a slight improvement.

Finally, the labels provided in some cases are rather formal and official terms that do not occur very frequently in the texts. For example, the term *buitenlandse werknemers* ('international employees') is less frequently used than the common term *gastarbeider* ('guest worker'). Similarly, in Dutch the term *acteur* is only used to denote male actors, while female actors are called *actrice*. For this reason, we expect further improvement if we take such synonyms and alternative terms into account. As a list of synonyms, the synonyms found by [23] were taken as a basis and manually corrected. The results of the runs using the synonym list are given in the third column of Table 1. In all cases the use of synonyms gives better results. The baseline benefits directly from the synonyms, while in the other cases the main effect is that more documents are taken into account to compute the co-occurrence distribution. The Novay algorithm was used in MediaEval 2010 [37] where it proved to be the stronger performer of the two algorithms used by participants in this task. The second algorithm [15], however, was interesting in its own right since it attempted to exploit a cluster-based approach.

4. TAGGING TASK, WILD WILD WEB

The *Tagging Task, Wild Wild Web* emulates the tagging activity of users in an online video community. Participants are required to predict the tags that uploaders assigned to Internet video. They are provided with a multilingual set of Creative Commons (CC) licensed Internet videos and the associated human-contributed metadata. The data set is accompanied by speech recognition transcripts in four languages [22]. Unique to the data set is that it is embedded within a social network consisting of user friendships and communications, although this information was not used by participants in 2010.

4.1 Use Scenario

The use scenario is the automatic generation of tags that were assigned by the users who uploaded the video to blip.tv (http://blip.tv). The task is motivated by the assumption that tags are interesting for search and browsing and that users would benefit from automatic methods that would predict tags and make sure that more videos had tags or additional tags. We do not differentiate between different kinds of tags, but try to predict every tag. Automatic generation could be applied as either fully automatic or a semi-automatic 'suggestion' process.

4.2 Data Set and Evaluation

This data set is a collection of CC-licensed Internet video, collected from blip.tv and created by the PetaMedia Network of Excellence (http://www.petamedia.eu). The data set was created in compliance with three main specifications. It had to be representative of Internet video, it had to be able to be made freely available and it had to be associated with a sufficiently dense social network. We noticed that users on Twitter publish tweets about videos. We decided to collect videos from shows for which we know that at least one episode of the show has been tweeted. We

used Topsy (http://topsy.com) to collect blip.tv links from tweets. Their licenses were checked to make sure that they were Creative Commons. The videos were downloaded from blip.tv. Then Topsy was searched again to gather all users that had mentioned any one of these videos. This set formed the Level 0 users. We collected up to 3200 posts from each user. Then we collected the list of users that they communicate with by directly sending them messages. These are Level 1 users. Then, we collected the profiles of Level 1 users and also of Level 2 users, i.e., the interlocuteurs of Level 1 users.

The data set was gathered from a range of blip.tv shows (i.e., channels). It contains ca. 350 hours worth of data for a total of 1974 episodes (247 development and 1727 test). The episodes were chosen from 460 different shows—shows with less than four episodes were not considered for inclusion in the data set. Only episodes for which the speech recognizer achieved an average word-level confidence score of > 0.7 were included in the set. The set is predominantly English with approximate 6 hours of non-English content divided over French, Spanish and Dutch.

Participants were provided with a video file for each episode along with metadata (e.g., title + description), speech recognition transcripts [22], shot-level information (shot boundaries plus one extracted keyframe per shot) [19] and social network information from Twitter described above. Note that the social network information was not used by any participants in 2010, but we describe it here since it is anticipated that it will be used by others who make use of the data set in the future. Note also that the speech recognition transcripts do not fall under the CC-license. They were kindly donated to MediaEval 2010 by LIMSI (http://www.limsi.fr) and Vocapia Research (http://www.vocapia.com) and are licensed separately from the rest of the data set for research use.

The ground truth consists of tags that have been assigned to the videos by users. We de-noised the tags, by choosing only high-frequency tags—tags occurring > 10 times in a large sample of blip.tv content. The result was a list of 746 tags for the development set and 1271 tags for the test set. The two sets of tags were not mutually exclusive.

The Tagging Task, Wild Wild Web is particularly challenging, not only because the videos are largely not recorded in professional studios and the spoken content is often spontaneously produced, but also because of the nature of the tags to be predicted. Users use tags designating abstract topics that should not be expected to appear in the spoken content of the videos (e.g., `world politics`)—this property of abstract topics was previously mentioned in Section 2. Also, tags that go beyond topic to aspects such as genre (e.g., `animation`) are also used. Finally, tags specific to particular people (e.g., `jim kirks`), series title (e.g., `the jama report`), personal taste judgments (e.g., `wow cool show`) or other aspects (e.g., `season 1`).

4.3 Algorithm and Results

In this section, we report on a the performance of a basic algorithm that approaches the task of assigning a subject label to a video as an information retrieval problem, treating the subject label as a query and the set of videos to be labeled as the collection. Like the Novay algorithm applied to the Tagging Task, Professional (cf. Section 3.3), it is an IR approach that makes use of the divergence model. The document model here, however, is less sophisticated. Instead, here the focus is set on comparing the performance between an approach using metadata only and an approach using speech recognition transcripts. The highest ranking videos returned by the system in response to the query, are assigned that query as a label.

Table 2: Performance on the *Tagging Task, Wild Wild Web* (Mean Average Precision)

	ASR	**metadata**	**ASR and metadata**
Development set	0.20	0.29	0.33
Test set	0.15	0.25	0.27

The specific challenge tackled by our algorithm is handling the shortness of the labels, which are one or at most two words long. This challenge is addressed by performing a round of pseudo-relevance feedback, i.e., expanding the query with important terms extracted from top-ranking documents returned by an initial retrieval round. This approach was shown to be effective during VideoCLEF 2008 [25]. The difference between that work and the work reported here is twofold. First, in [25] the Vector Space Model is used, whereas here, we apply the more recent language modeling framework for information retrieval. Second, the MediaEval data set is over ten times as large as the VideoCLEF 2008 test set and contains more challenging, user-generated material. Pseudo-relevance feedback can lose its effectiveness in the face of noise in the data sets and helpful documents can be in danger of getting lost. For this reason, application of this approach to the Tagging Task, Wild Wild Web task provides an interesting, but non-trivial, baseline. In each case, we apply one round of pseudo-relevance feedback. For retrieval we use the Kullback-Leibler divergence model. To compute the Retrieval Status Value (RSV), this algorithm uses the negative divergence between multinomial models of the query and the document:

$$RSV = -D(\theta_Q \| \theta_D) = -\sum_w p(w|\theta_Q) \log \frac{p(w|\theta_Q)}{p(w|\theta_D)} \quad (4)$$

Here w is a word and θ_D and θ_Q are models of the document and the query. The sum is taken over all the words in the vocabulary. The smoothing method we choose is Bayesian smoothing with a Dirichlet prior.

Results are reported in Table 2 in terms of Mean Average Precision (MAP). The results support our starting assumption that indexing with metadata only would outperform indexing with ASR transcripts and that the optimal performance is obtained when both sources are used together.

Participants in MediaEval 2010 made use of either speech recognition transcripts or of human-contributed metadata in order to approach the task. The IR approach, chosen by [13], in general outperformed the classification approach, chosen by [3]. However, [3] was able to achieve a significant boost in performance by making use of the file names of the videos, which contained information on the identity of the uploader. Also, [3] was the only MediaEval 2010 participant to address the open-set tagging task in additional to the closed-set tagging task. An interesting result was that using only words in the recognizer transcripts with high score did *not* improve performance over using all words in the speech

recognition transcript [13]. In the future, it would be interesting to experiment on a range of different content with different speech recognition error levels, rather than considering only content that achieves a score of at least 0.7.

5. PLACING TASK

The *Placing Task*, requires participants to automatically assign geo-coordinates (i.e., geotags) to videos from Flickr. Only about 4% of images at Flickr are geotagged. Recently, Flickr allowed users to share videos of up to 90 second in length. Geo-coordinates are often associated directly with images at the moment that they are captured by the camera. Videos differ from images because they usually must be geotagged manually by the user, using a map interface, and far fewer videos than images have been geotagged.

5.1 Use Scenario

Users would like to place personal videos on a map without any significant manual effort. The geotagging system suggests the most probable geographic coordinates for an uploaded video, based on visual features extracted from its frames and based on user contributed metadata (e.g., title, description and tags). Any other information, like permanent user locations or tags of previously uploaded videos and images, also might contain clues about where the video should be placed on a map.

5.2 Data Set and Evaluation

The MediaEval 2010 Placing Task data set consists of CC-licensed videos that were crawled from Flickr. Videos are in mp4 format and include the Flickr metadata. The metadata for each video includes user-contributed title, tags, description, comments and also information about the user who uploaded the videos. Information about the user's contacts, favorites, and all videos uploaded in the past are also included. The data set is divided into training data (5091 videos) and test data (5125 videos). Videos were selected both to provide a broad coverage of users, and also because they were geotagged with a high accuracy at the 'street level'. Accuracy shows the zoom level the user used when placing the photo on the map. There are 16 zoom levels, and these correspond to 16 accuracy levels (e.g., 6—region level, 12—city level, 16—street level). The sets of users from the test and the training collections were disjoint, to allow for the most challenging cold-start scenario. For development purposes the dataset also contained metadata extracted from a large set of Flickr images. Using geographic bounding boxes of various sizes and the Flickr API, the metadata for 3,185,258 CC-licensed Flickr photos were collected by uniformly sampling from all parts of the world. Most, but not all, photos have textual tags. All photos have geotags of at least region-level accuracy. The dataset also contains visual features extracted for both photos and frames of the videos (a frame at every fourth second of video was extracted and saved in jpeg format). Nine visual features were extracted using the open source LIRE library (http://www.semanticmetadata.net/lire) with the default parameter settings and the default image size which is 500 pixels on the long side. Features included the following descriptors: Color and Edge Directivity Descriptor, Gabor, Fuzzy Color and Texture Histogram, Color Histogram, Scalable Color, Auto Color Correlogram, Tamura, Edge Histogram, and Color Layout. Evaluation was done by calculating the distance from the actual point (assigned by a Flickr user, owner of the video) to the predicted point (generated automatically by an algorithm). While it was important to minimize the distances over all test videos, runs were compared by finding how many videos were placed within a threshold distance of 1 km, 5 km, 10 km, 50 km and 100 km.

The participants in the MediaEval 2010 Placing Task made use of geotagging methods that fall roughly into three categories. First, the most basic and most popular method, uses the presence of location names in the video metadata. Terms extracted from the metadata are mapped to geo-coordinates using a gazetteer. A popular choice of gazetteer is GeoNames (http://www.geonames.org). Application of this method typically involves resolution of ambiguous place names and also pre-processing that compensates for variation in the language, form or spelling of location names used by the users who create the metadata. Second, geo-coordinates are predicted by finding similar items in the training set and propagating their geo-coordinates to a test item. Third, training items are divided down into geographical regions, using either clustering or a fixed-size grid and a model is estimated for each region. Such models can be build using any available features, with obvious choices being visual features and textual metadata. Metadata derived models leverage the fact that some object names (for example 'double-decker bus') and person names ('Aung San Suu Kyi') suggest locations, although they are not themselves location mentions. The system that achieved the best precision at MediaEval 2010 (submitted by the authors of [36]) made use of a combination of the second and third method, first narrowing the location of a video using a model and then assigning exact coordinates by identifying its nearest neighbors within that location.

5.3 Algorithm and Results

The algorithm uses an analysis of the user-contributed tags of a video in order to predict that video's geotag. It was developed by ICSI (http://www.icsi.berkeley.edu), one of the teams that participated in the MediaEval 2010 Placing Task [5]. Other papers addressing the Placing Task at MediaEval 2010 include [36, 20]. The basic strategy underlying the algorithm is to assign the video the geo-coordinates that are associated with the tag that is determined to be most closely associated with a well-defined spatial area on the map. The approach makes use of the prior distribution of tags in a training set of previously tagged resources (which can include videos and images) for which the geotags are known. Exploratory experiments demonstrated that this approach outperforms a method based on automatically extracting location-related words from the video metadata and performing gazetteer lookup.

The tag-based approach is motivated by the following considerations. It was observed that it is quite challenging to exploit visual features, since they sometimes have low correlation with the location where a video is recorded (e.g., indoor scenes). On the other hand, Flickr video is richly endowed with user-contributed metadata. Titles, tags, and descriptions contributed by the user often provide direct and sensible clues for the task of location prediction. Of the videos in the 2010 Placing Task training set, 98.8% have at least a title, tag, or description in their metadata. At least one tag is associated with 88.1% of them.

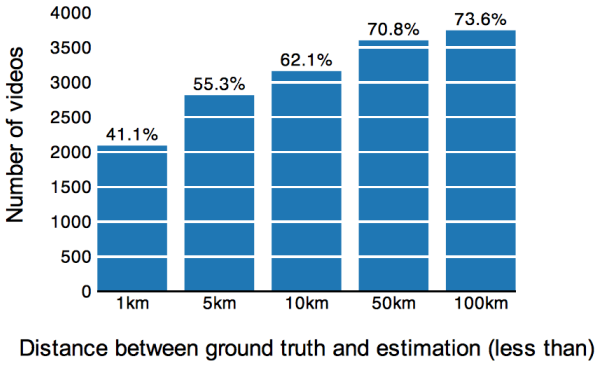

Figure 1: Performance on the *Placing Task* test set using the prior distribution of tags and limiting the search scope to the same user's uploads.

In order to find the tag that is most closely associated with a well-defined spatial area, the ICSI algorithm uses a data-driven approach that exploits the generalization that the geographical relevance of a given tag will be related to spatial distribution of resources assigned that tag. For each tag associated with the test video, all training resources with that tag are plotted in the 2D coordinate plane. Since the training images do not contain title or description in their metadata, only tags were used for this experiment. The co-ordinates associated with the smallest spatial variance are predicted as the co-ordinates of the test video. The spatial variance is calculated by counting the number of videos within a region of a given radius and normalizing by the total number of videos with that tag. The video is assigned the geo-coordinates of the center of the region defined by the tag determined to have the smallest spatial variance. The normalization has the purpose of controlling the influence of more frequent but less spatially significant tags, such as `video` or `2009`. It was noticed that the algorithm quite often returned the tag corresponding to a toponym representing the smallest geographical entity. Even in cases where the selected tag was not a toponym, it still picked a useful point from the area of highest concentration in the tag region.

The user ID of the uploader, which is contained in the metadata, proved valuable in refining the improvement of the performance of this approach. The best performance was achieved when geo-coordinates were assigned to a given video based only on the spatial distribution of other videos uploaded by the same user. Since each person has a idiosyncratic method or personal style for choosing a tag for certain events, this scheme has an improved chance of finding geographically related videos. The identity of the uploader was only used in the case that other resources from the same uploader existed in the collection, As can be seen in Figure 1, the system estimated locations for 41.1 % of videos within the 1 km range of ground truth, and 73.6 % within the 100 km range of ground truth.

6. CONCLUSION AND OUTLOOK

Automatically generated tags and geotags can supplement user-contributed annotations and contribute to improving access to content in video collections and communities. We presented three tasks involving tagging and geotagging that were run in MediaEval 2010 along with descriptions of their data sets, definitions and a reference algorithm. Algorithms developed to approach these tasks can profitably make use of language-based features (e.g., derived from user-contributed metadata), spoken content and also context. The focus on these three sources of information sets MediaEval apart from other video retrieval benchmarks. We observed several general trends in MediaEval 2010. First, user-contributed or human-generated metadata, if available, makes an important contribution to tagging and geotagging. Second, features derived from the spoken audio or visual channel also have a contribution to make, although they are generally more difficult to exploit. Statistics that exploit collection level information (i.e., co-occurrences of words, co-location of tags) can be used to improve performance. External resources are also important. MediaEval 2011 will again offer tagging and geotagging tasks, on a larger scale and formulated to increase the level of challenge, while keeping tasks close to the original real-world use scenarios that motivated them. In particular, we intend to offer a tagging task that promotes the development of systems that tackle the prediction of different types of tags in different manner. We will foster the strengthening of the use of multiple modalities—both by encouraging combination and by strengthening individual modalities (e.g., speech) in isolation. Finally, we will continue to encourage participants to specifically address those aspects of the task that are specific to video. In particular, the temporal structure of video, which sets video apart from still images, will become increasingly important.

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from EC FP7 under grant agreement no 216444 (PetaMedia Network of Excellence). We would also like to thank the Glocal project (grant agreement no 248984) for their role in organizing the 2010 Placing Task. We also acknowledge the contributions of Jaeyoung Choi and Adam Janin who, along with Gerald Friedland, developed the ICSI location estimation algorithm. Finally, we would like to express our appreciation to all the researchers who participated in MediaEval 2010 and assisted with the organization.

8. ADDITIONAL AUTHORS

Stevan Rudinac (Delft Univ. of Technology, email: s.rudinac@tudelft.nl), Christian Wartena (Novay, email: christian.wartena@novay.nl), Vanessa Murdock (Yahoo! Research, email: vmurdock@yahoo-inc.com), Gerald Friedland (International Computer Science Institute, email: fractor@icsi.berkeley.edu), Roeland Ordelman (Netherlands Institute for Sound and Vision and Univ. of Twente, email: rordelman@beeldengeluid.nl) and Gareth J.F Jones (Dublin City Univ., email: gareth.jones@computing.dcu.ie)

9. REFERENCES

[1] J. Allan, editor. *Topic Detection and Tracking: Event-based Information Organization*. Kluwer, 2002.

[2] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: Geotagging web content. In *SIGIR '04*, pages 273–280, 2004.

[3] K. Chandramouli, T. Kliegr, T. Piatrik, and E. Izquierdo. QMUL @ MediaEval 2010 Tagging Task: Semantic query expansion for predicting user tags. In *MediaEval '10 Working Notes*, 2010.

[4] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: A content-based approach to geo-locating Twitter users. In *CIKM '10*, pages 759–768, 2010.

[5] J. Choi, A. Janin, and G. Friedland. The 2010 ICSI video location estimation system. In *MediaEval '10 Working Notes*, 2010.

[6] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.

[7] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *WWW '09*, pages 761–770, 2009.

[8] C. Danesi and C. Clavel. Impact of spontaneous speech features on business concept detection: A study of call-centre data. In *SSCS '10*, pages 11–14, 2010.

[9] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *CIKM '98*, pages 148–155, 1998.

[10] F. Eisterlehner, A. Hotho, and R. Jäschke, editors. *Proceedings of the ECML PKDD Discovery Challenge 2009*, Sept. 2009.

[11] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning. Domain-specific keyphrase extraction. In *IJCAI '99*, pages 668–673, 1999.

[12] J. Garofolo, G. Auzanne, and E. Voorhees. The TREC Spoken Document Retrieval Track: A success story. In *RIAO '00*, pages 1–20, 2000.

[13] A. Gyarmati and G. J. F. Jones. DCU at MediaEval 2010—Tagging Task Wild Wild Web. In *MediaEval '10 Working Notes*, 2010.

[14] J. Hays and A. Efros. IM2GPS: Estimating geographic information from a single image. In *CVPR '08*, pages 1–8, 2008.

[15] D. Hernández-Aranda, R. Granados, J. Cigarran, A. Rodrigo, V. Fresno, and A. García-Serrano. UNED at MediaEval 2010: Exploiting text metadata for automatic video tagging. In *MediaEval '10 Working Notes*, 2010.

[16] P. Heymann, D. Ramage, and H. Garcia-Molina. Social tag prediction. In *SIGIR '08*, pages 531–538, 2008.

[17] A. Hotho, B. Krause, D. Benz, and R. Jäschke, editors. *Proceedings of the ECML PKDD Discovery Challenge 2008*, Sept. 2008.

[18] M. Huijbregts, R. Ordelman, and F. de Jong. Annotation of heterogeneous multimedia content using automatic speech recognition. In *Semantic Multimedia*, volume 4816 of *LNCS*, pages 78–90. Springer, 2007.

[19] P. Kelm, S. Schmiedeke, and T. Sikora. Feature-based video key frame extraction for low quality video sequences. In *WIAMIS '09*, pages 25–28, 2009.

[20] P. Kelm, S. Schmiedeke, and T. Sikora. Multi-modal, multi-resource methods for placing Flickr videos on the map. In *ICMR '11 (this proceedings)*, 2011.

[21] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *SIGIR '01*, pages 111–119, 2001.

[22] L. Lamel and J.-L. Gauvain. Speech processing for audio indexing. In *Advances in Natural Language Processing*, volume 5221 of *LNCS*, pages 4–15. Springer, 2008.

[23] V. Malaisé, A. Isaac, L. Gazendam, and H. Brugman. Anchoring Dutch Cultural Heritage Thesauri to WordNet: Two Case Studies. In *LaTeCH '07*, pages 57–64, 2007.

[24] M. Montagnuolo and A. Messina. Parallel neural networks for multimodal video genre classification. *Multimedia Tools Appl.*, 41(1):125–159, 2009.

[25] E. Newman and G. J. F. Jones. DCU at VideoClef 2008. In *Evaluating Systems for Multilingual and Multimodal Information Access*, volume 5706 of *LNCS*, pages 923–926. Springer, 2009.

[26] J. S. Olsson and D. W. Oard. Improving text classification for oral history archives with temporal domain knowledge. In *SIGIR '07*, pages 623–630, 2007.

[27] G. Paaß, E. Leopold, M. Larson, J. Kindermann, and S. Eickeler. SVM classification using sequences of phonemes and syllables. In *Principles of Data Mining and Knowledge Discovery*, volume 2431 of *LNCS*, pages 373–384. Springer, 2002.

[28] P. Pecina, P. Hoffmannová, G. J. F. Jones, Y. Zhang, and D. W. Oard. Overview of the CLEF 2007 Cross-Language Speech Retrieval Track. In *Advances in Multilingual and Multimodal Information Retrieval*, volume 5152 of *LNCS*, pages 674–686. Springer, 2008.

[29] T. Rattenbury and M. Naaman. Methods for extracting place semantics from Flickr tags. *ACM Transactions on the Web*, 3(1):1, 2009.

[30] S. Rendle and L. Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *WSDM '10*, pages 81–90, 2010.

[31] R. C. Rose, E. I. Chang, and R. P. Lippmann. Techniques for information retrieval from voice messages. In *ICASSP '91*, pages 317–320, 1991.

[32] P. Serdyukov, V. Murdock, and R. van Zwol. Placing Flickr photos on a map. In *SIGIR '09*, pages 484–491, 2009.

[33] B. Sigurbjörnsson and R. Van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW '08*, pages 327–336, 2008.

[34] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, 2006.

[35] A. Stolcke et al. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373, 2000.

[36] O. Van Laere, S. Schockaert, and B. Dhoedt. Finding locations of Flickr resources using language models and similarity search. In *ICMR '11 (this proceedings)*, 2011.

[37] C. Wartena. Using a divergence model for MediaEval's Tagging Task (Professional Version). In *MediaEval '10 Working Notes*, 2010.

[38] C. Wartena and R. Brussee. Topic detection by clustering keywords. In *DEXA '08*, pages 54–58, 2008.

[39] C. Wartena, R. Brussee, and W. Slakhorst. Keyword extraction using word co-occurrence. In *DEXA '10*, pages 54–58, 2010.

[40] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM '01*, pages 403–410, 2001.