Blip 10000 : A social Video Dataset containing SPUG Content for Tagging and Retrieval

Sebastian Schmiedeke, Peng Xu, Isabelle Ferrané, Maria Eskevich, Christoph Kofler, Martha Larson, Yannick Estève, Lori Lamel, Gareth Jones, Thomas Sikora

Blip10000: A social Video Dataset containing SPUG Content for Tagging and Retrieval

Sebastian Schmiedeke[1] Peng Xu[2] Isabelle Ferrané[3]
Maria Eskevich[4] Christoph Kofler[2] Martha A. Larson[2]
Yannick Estève[5] Lori Lamel[6] Gareth J.F. Jones[4] Thomas Sikora[1]
[1]Technische Universität Berlin, Germany [2]Delft University of Technology, The Netherlands
[3]University of Toulouse, France [4]Dublin City University, Ireland,
[5]Language and Speech Technology (LST) team, LIUM, Le Mans, France
[6]Spoken Language Processing Group, LIMSI/Vocapia, France

ABSTRACT

The increasing amount of digital multimedia content available is inspiring potential new types of user interaction with video data. Users want to easily find the content by searching and browsing. For this reason, techniques are needed that allow automatic categorisation, searching the content and linking to related information. In this work, we present a dataset that contains comprehensive semi-professional user-generated (SPUG) content, including audiovisual content, user-contributed metadata, automatic speech recognition transcripts, automatic shot boundary files, and social information for multiple 'social levels'. We describe the principal characteristics of this dataset and present results that have been achieved on different tasks.

Keywords

Dataset, SPUG Content, Video Tagging, Speech Retrieval

1. INTRODUCTION

The explosive growth of the amount of internet videos available online creates continuing challenges and opportunities for the video indexing and retrieval techniques. Extensive, comprehensive and publicly internet video datasets can not only provide a platform to develop and validate new algorithms, but can also be a valuable resource that can be used to analyse the user behaviour. Exciting and novel problems can be defined on the basis of a large representative dataset composed of contributions from many users.

The requirements for a well-designed internet video dataset involves several aspects. First, the data collection strategy should be well designed to ensure sufficient scale and content diversity. In this way, the dataset can represent internet videos in a manner that is as unbiased as possible. Second, internet videos are naturally associated with multi-modal information, this information should be collected comprehensively and be well organized, in order to analyse the contribution of specific information resource in a certain task. In particular, the development of social networks provide rich contextual information for internet videos. The social information has proven to be valuable in many information retrieval domains [9], demonstrating its exciting potential for internet video analysis. Third, the definition of the tasks and generation of the ground truth should be based on real-world user scenarios and also be appropriate for metric-based evaluation, so that the methods derived based on the dataset can be comparable to each other and over time.

Motivated by these concerns, we present the Blip10000 dataset, which consists of 14,838 videos for a total of 3,288 hours from `blip.tv`. During dataset generation we invested effort in including a combination of combination of information from audiovisual content, user-contributed metadata, automatic speech recognition (ASR) transcripts and social networks. The videos cover a broad range of topics and styles. To facilitate the usage of the dataset by researchers from different research communities, we provided the ASR transcripts as well as the shot boundary detection results based on the state-of-the-art algorithms. The social information has been particularly emphasized in the data collection stage. Using the search engine Topsy, we exploit the connection between the videos and the information from Twitter. The social network contextual information includes user profiles and the tweets associated with the videos. This information has great potentials for improving various video applications, but has not yet been fully exploited for video indexing and retrieval.

The Blip10000 dataset, or its subset has been used used by the MediaEval Multimedia Benchmarking[1] tasks from 2010 to 2012. As a subset of the Blip10000, ME10WWW (1,974 videos) was used in the 2010 Wild Wild Web (WWW) tagging, 2011 Genre Tagging and Rich Speech Retrieval tasks. The entire Blip10000 dataset is used in the 2012 Genre Tagging task, the Search and Hyperlinking task. The definition of tasks focus on modelling the real world user scenario. For

[1]`http://www.multimediaeval.org`

example, in the Genre Tagging task, each video is supposed to be assigned one of the genre related tags, which are defined by `blip.tv`. In the Rich Speech Retrieval task and Search and Linking task, the crowdsourcing web site Amazon Mechanical Turk (MTurk) is exploit to generate queries and ground truth labels. With the help of this dataset, people have developed effective techniques for these tasks. On the other hand, through these tasks, people can get further understanding of this dataset as well as the represented web videos archives.

Researchers are more than welcome to create other tasks based on the proposed dataset. It can either be user-centric tasks with application specific targets, or data-driven problems aiming to investigate the nature of internet videos. Note that every dataset is a subset of all the videos on the web. The investigation of the different behaviours between Blip10000 and other alternative datasets can be valuable contributions.

1.1 Related Work

Many internet video datasets developed for research purposes are limited to certain domain applications, such as action recognition [11] or near-duplicate detection [17]. Only a few datasets have been proposed for web video indexing or retrieval.

Zanetti [19] collected a YouTube video collection for categorization purpose. The ground truth categorizes are defined based on the YouTube categories. This dataset mainly focus on the audiovisual content of web videos, so the information from other domain is quite limited.

A recent dataset Columbia Consumer Video Database [4] contains more than 9000 videos with event related categories. MTurk is also used for annotation. Although consumer video is a broad domain with very diverse content, they are different from other web videos in the fact that consumer videos have much less textual information. For this reason, the work based on this dataset also mainly considers audiovisual content.

MSRA-MM dataset [8] is a comprehensive web video dataset with more than 20K videos and human label ground truth. The disadvantage of this dataset lies in the copyright issues. Only the links to the web page can be provided and not the video data. The textural information associated with certain video is indexed together, so that it is difficult to analyse the specific contribution of various text information resources. Since all the content of Blip10000 dataset are with Creative Comments lessons, it is much more feasible for user to access.

TRECVid [10], is a large video retrieval benchmark in multimedia community. Early on, only professional edit TV news or other TV programs are used in TRECVid. From 2010, internet videos have been used in several tasks. The design of TRECVid tasks is mainly focused on exploiting visual information for applications on the shot level (concept detection), or short video clips (event detection). In contrast, Blip10000 is associated with tasks which are interested with the overall meaning of entire video. The aim of Blip10000 is to support emerging techniques that exploit multi-modal information resources, especially the potential of social information.

2. BLIP10000 DATASET

The Blip10000 dataset contains videos from `blip.tv`. The `blip.tv` content is created by users who have gone beyond the point-and-shoot video capture methods common on platforms such as YouTube and Flickr. `Blip.tv` contributors demonstrate at least basic proficiency in filmmaking. Such content is generally referred to as semi-professional user-generated (SPUG) content, which tends to be scripted or well thought out. In general, it is aimed specifically at communicating a message or opinion or at entertainment. `Blip.tv` users publish video content in a series to one particular topic, publishing at regular intervals, and targeting a broad audience. These shows cover a range of topics and styles.

2.1 Querying Twitter for Video Links

The videos were collected for shows for which the link to one of their episodes has been mentioned in Twitter messages of users tweeting about them. For this reason, Topsy[2] was used to collect links to `blip.tv` videos from these tweets. After a few rounds of crawling this search engine, a list of 25,005 tweets mentioning `blip.tv` links was received from 8,814 users. These tweets are dated from Nov. 2007 to Nov. 2009. The motivation for posting links is quite different; some posts comment an episode(i. e. **FANTASTIC* keynote presentation from @timoreilly [...] Work on important things! Get involved! Do something!*), while other posts just announce an upload (i. e. *Posted 'Vlog - I Caught A Case Of The Mondays' to blip.tv [...]*). Then, the license and availability of the videos behind these discovered `blip.tv` links were checked. This ended up in 2,237 different `blip.tv` uploaders, from whose shows all available episodes were taken. All downloaded videos were checked that they were licensed under Creative Commons. The whole dataset contains 14,838 episodes comprising a total of ca. 3,288 hours of data, so an episode lasts 13 minutes in average. The partition into development set and test set is described in section 2.2.

Each video is associated with metadata (e.g., title, description, tags, ID of uploader), social network information (i. e. , Twitter messages), automatic speech recognition (ASR) transcripts, and shot boundary information including a key frame per shot. Each video is associated with only one genre/ category label[3]. The following sections describe these different parts of the dataset in more detail.

2.2 Creating Sets

This dataset was designed for tagging and retrieval tasks. To facilitate work in these areas, the episodes were separated into a development set and a test set. The development sets contains a third of the runtime and the episodes number of the whole data set—these 5,288 videos having a runtime of 1,143 hours. Respectively, the other 9,550 videos (with a runtime of 2,145 hours) belong to the test set. Concerning

[2] `http://topsy.com`

[3] `Art`, `Autos and Vehicles`, `Business`, `Citizen Journalism`, `Comedy`, `Conferences and Other Events`, `Documentary`, `Educational`, `Food and Drink`, `Gaming`, `Health`, `Literature`, `Movies and Television`, `Music and Entertainment`, `Personal or Auto-biographical`, `Politics`, `Religion`, `School and Education`, `Sports`, `Technology`, `The Environment`, `The Mainstream Media`, `Travel`, `Videoblogging`, `Web Development and Sites`, `Default Category`.

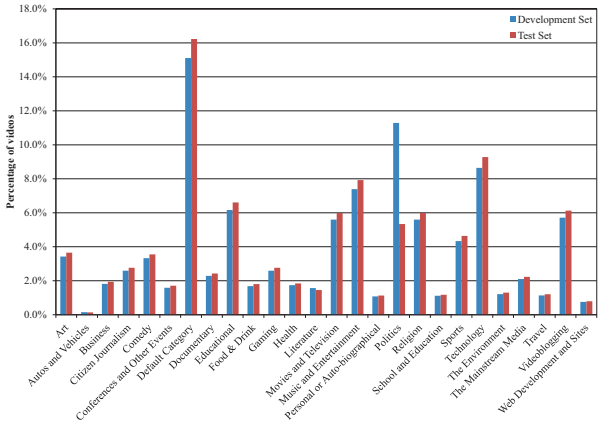

Figure 1: Genre distribution among both sets.

this partition and the size of the sets, we enable the direct application of both retrieval and classification approaches to address our proposed tasks. The genres are distributed among both sets accordingly to this propotion; one third in the development set, and two third in the test set. The show id or uploader id was not taken into account for this separation. So, these sets do not mutual exclude shows (and therefore its uploaders) or it is not ensured that all shows occur in both sets. In fact, there are 803 shows its episodes occur in both sets, while the total number of shows for the development set is 1220, and 1820 for the test set, respectively.

As shown in Figure 1, a few categories dominate the dataset that have implications for the training of appropriate classifiers and the choice for evaluation metrics. The majority of episodes is assigned to one of the following five genre categories: 'Technology', 'Music and Entertainment', 'Politics', 'Educational', or 'Default Category'. It should be noticed that episodes associated with the default category may topically belong to another one. Figure 1 also shows that the desired category distribution of one third could not be achieved for the 'Politics' genre, since all videos from the subset ME10WWW [7] previously offered should be merged into the development set.

2.3 Video Content and Labels

The downloaded videos were converted into the container format `ogg` that is unrestricted by software patents using Theora as video codec and Vorbis as audio codec, respectively. The conversion was performed with the constraint that the original audiovisual characteristics like resolution, frame rate, sample rate, and audio channels, were persevered. So, the data set contains video files with heterogeneous specifications (i. e. , the resolutions occurred vary from CIF to FullHD).

For each episode, shot boundaries were provided by TU Berlin. This shot segmentation was carried out automatically by a software implementation [5]. Note that because of the automatic detection procedure shot boundary information will not necessarily be perfect. For each shot segment, a key frame is extracted from the middle of the shot. In total, this dataset includes approximately 420,000 shots/ key frames, concluding an average shot length of about 30 seconds. Each video is associated with exactly one of the 26 already mentioned genre labels. These genre labels were determined by querying the `blip.tv` web API[4]. The genre label of each video is represented by the field `categoryName` in the `JSON` output provided by the API. Subsequently, the genre labels were normalized by replacing whitespaces with underscores ('_') and ampersands ('&') by the word 'and'. The category of each episode was chosen by its uploader. In case that not any category was chosen, these episodes were assigned to the default category. To avoid rare genre categories, episodes were merged to the default category if they were associated to categories that have less than 100 assigned episodes. We made one exception for episodes assigned to the category 'Autos and Vehicles'—containing only 21 episodes—to be consistent with a subset previously offered (ME10WWW [7]). The ground truth data is provided as plain text file that is compatible with the trec_eval software[5].

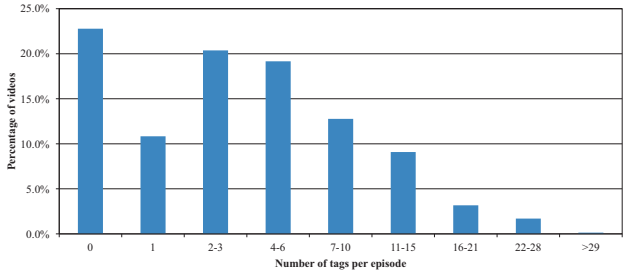

Figure 2: Distribution of number of tags per video.

2.4 User-contributed Metadata

The metadata consists of the information that was assigned by the creator to the `blip.tv` episode upon upload. These data is stored in UTF-8-formatted XML files for each video including information about the *title, description, uploader id, license, duration,* and *tags*. In particular, the <title> (contains the episode title) element and the <description>(contains a description of the episode) element in this metadata can be useful for tags propagation or genre prediction.

The titles and descriptions taken from `blip.tv` are preserved using CDATA sections, so these section can contain markup elements originally occurring. Since we want to represent a naturally existing multimedia collection and we want to preserve its 'wild' character, our collection subsequently contains metadata in various languages. The language is predominantly English, but also non-English content (i. e. , in French, Spanish, German or Dutch) occurs. Concerning this fact, we performed a normalization to the tags: They are formatted to be in lower case and they should not contain any special characters, like diacritics, symbols, or punctuation—including whitespaces.

This data set contains 2034 unique tags, its most frequent tag ('obama') occurs in 3.4 % of all episodes. The Figure 2 depicts how users tag their episodes. The majority of videos are tagged with less than four key word; among these, the largest partition is the one containing episodes that have no tag.

[4] `http://wiki.blip.tv/`
[5] `http://trec.nist.gov/trec_eval/`

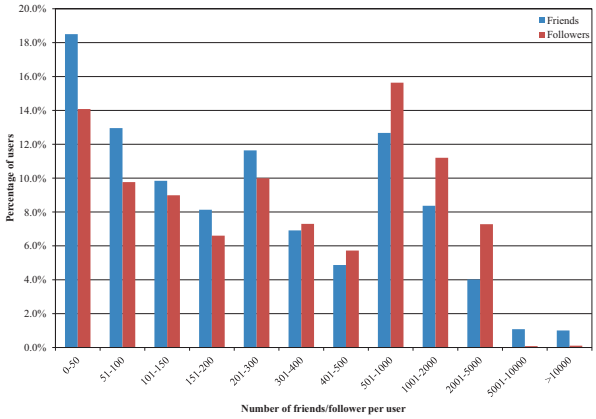

Figure 3: Histogram of number of friends/followers per user.

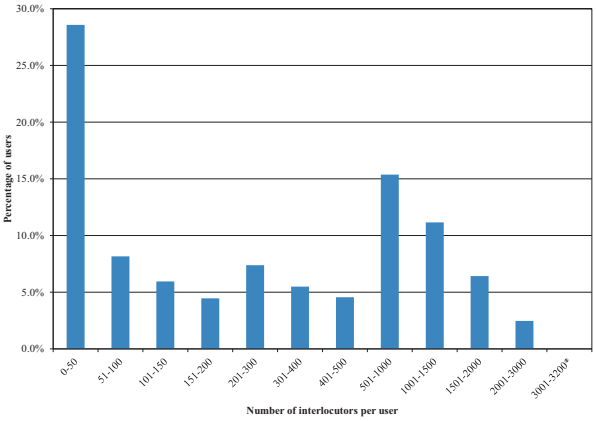

Figure 4: Histogram of number of direct communications per user. *Maximum number of 3200 retrieved direct replies is limited by Twitter API.

2.5 Social Net

Since we collected only videos from shows on `blip.tv` for which we knew that at least one episode from that show had been mentioned in a tweet, the social data was also gathered from Twitter.Twitter was a good choice because we were able to easily establish the connection between `blip.tv` and tweets via this real-time search engine `Topsy`.

First, we search Topsy for all Twitter users who mentioned a particular episode in their tweets, resulting in 8,814 unique users mentioning videos contained in this dataset. The subset of accessible profiles (`8,436`) presents the '0th social level'. Based on these Twitter users, we then used the Twitter API for crawling seed user profiles using a white-listed IP address. Each seed user's profile includes the list of his 'friends' (persons whom he is following), his followers (persons who are following him), and his interlocutors ('@') among with personal data such as *name*, *nick name*, *location* and a short *statement*. Figure 3 depicts the ratio between friends and followers of the seed users. The total number of unique friends is 1,513,716 within the 0th level.

Figure 4 shows the distribution of interlocutors per Twitter user within the 0th level. The '1st social level' is constituted by each author's interlocutors and includes `410,380` accessible profiles (421,347 in total). These contacts' own friends constitute the '2nd social level' that comprise `9,106,467` Twitter users whose profiles are not provided.

The number of profiles provided at each 'social level' is listed in bold print in Table 1. It should be noticed that the actual number of provided profiles is lower due to publicly inaccessible user information. The users at each level were crawled based on the seed users at the previous level.

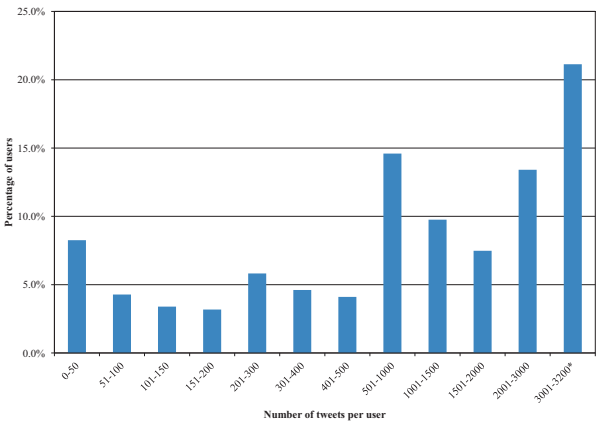

Figure 5: Histogram of number of tweets per user. *Maximum number of 3200 retrieved tweets is limited by Twitter API.

Table 1: Unique Twitter user per social level; *seed users of the corresponding level.

	Level 0	Level 1	Level 2
seed	**8,814***		
interlocutors		**421,347***	
friends		1,513,716	9,106,467
follower		1,513,716	

Twitter was not only used to create connections between uploaders and viewers, but also to have potentially annotations and comments about the `blip.tv` episodes. Therefore, up to 3,200 latest posts were crawled per '0th level' user, this is the limit of the Twitter API. The activeness of these user in terms of posting tweets is shown in Figure 5. Among these tweets, `blip.tv` episodes were mentioned 25,005 times.

2.6 Speech Transcripts

The speech transcripts were extracted from the audio streams that were preprocessed using a combination of `ffmpeg` and `sox` software. The audio streams were downmixed to a mono channel and downsampled to a sample rate of 16 kHz. The automatic speech recognition transcripts were generously provided by LIMSI/Vocapia[6] and LIUM Research team (LST)[7]. The data is predominantly English, but there are also small numbers of Czech, Dutch, French, Italian, German and Spanish shows present. Depending on the source (LIMSI/Vocapia or LIUM), the transcripts are accompanied by sets of complementary information or scores.

[6] `http://www.vocapia.com/news/2011_07_15.html`
[7] `http://www-lium.univ-lemans.fr/en/content/language-and-speech-technology-lst`

2.6.1 LIMSI/Vocapia

LIMSI/Vocapia [6] provided an XML file for each audio file 'successfully' processed—that are 5,237 files for the development set and 7,215 for the test set, respectively. Transcripts that include alternate hypotheses obtained from a consensus network were produced for all the above languages, according to the following strategy. The language identification detector (LID) automatically identified the language spoken in the whole video along with a language confidence score (`lconf`). However the LID results were not manually checked. Each file with a language identification score equal or greater than 0.8 was transcribed with the detected language. The remaining files were transcribed twice, with the detected language as well as with the English system. The average word confidence scores (`tconf`) were compared and the transcription with the higher score was chosen. There were files with other identified language for which there was no transcription system. In such cases, no transcripts were provided and for the remaining files no speech was detected.

2.6.2 LIUM Research team (LST)

The LIUM system [12] was developed to participate to the evaluation campaign of the international Workshop on Spoken Language Translation 2011. LST provided an English transcription for each audio file 'successfully' processed, that is 5,084 from the development set and 6,879 from the test set. These results consist of: (1) one-best hypotheses under NIST CTM format, (2) word-lattices under HTK Standard Lattice Format (SLF), following a 4-gram topology, and (3) confusion networks, under a ATT FSM-like format.

3. EVALUATIONS BASED ON BLIP10000

One major motivations of MediaEval has been to emphasise the 'multi' in multimedia and focus on human and social aspects of multimedia tasks. As previously mention, the Blip10000 dataset was designed to enable the use of features derived from audio (A), speech (S) and/or visual (V) content as well as from associated textual metadata (M) or social information.

3.1 The Tagging Task

This task aims to automatically assign genre labels to SPUG videos using different methods and sets of features. Genre is understood as related to common browsing categories used for Internet video sharing websites, in particular, by `blip.tv`. A similar task was evaluated in 2011 but only on the smaller subset ME10WWW.

The ground truth is provided by the genre label, based on the 26-genre list described in section 2.3, and which was associated to a video by its uploader. To perform a tagging task predicting category labels, we recommended the mean average precision (mAP) metric due to the biased category distribution.

Participants were encouraged to submit results corresponding to different systems. Five categories of runs were proposed depending on the media the features were derived from: audio and/or visual data, ASR transcripts, all data except metadata, all data except the uploader id or all data without restrictions.

This year, 6 teams have successfully participated in this task and have described their approaches in their respective publication: ARF [3], KIT [15], TUB [13], TUD [16], TUD_MM [18], and UniCamp [1]. Table 2 shows the best mAP value reached by each participant's best system regarding the feature domain used. This can be compared to the baseline results respectively obtained when all videos are assigned to the default category (mAP=0.0063), and randomly assigned (mAP=0.002). The size of the development set enabled participants to train models and develop systems based on classification approaches.

Table 2: Participants' best results within MediaEval Tagging Task for different feature domains.

	ARF	KIT	TUB	TUD	TUD_MM	UniCamp
A	0.1892					
V		0.3581	0.2301			0.1238
S	0.2174		0.1035	0.2536	0.3127	0.0027
M			0.5225			0.2112
A+V	0.1941					
A+V+S	0.2204					
V+S					0.2279	0.1238
S+M	0.3793					
V+S+M					0.3675	
Social+V			0.3431			

3.2 The Search and Hyperlinking Task

The Search and Hyperlinking task at MediaEval 2012 was a follow up on the experiments on search and linking of the multimedia data previously carried out in MediaEval 2011 Rich Speech Retrieval (RSR) task (used ME10WWW) and VideoCLEF 2009 Linking Task. The novelty of the task consists in the combination of search and linking within one framework, and the use of this larger dataset. Although the task contains one scenario, it is divided into 2 sub-taks: to retrieve video segments corresponding to textual or multimedia queries, and to use either search sub-task ground truth videos or search sub-task output results as anchor videos in order to form links to other videos in the collection that enrich user search experience.

Search sub-task methods focused on use of different ASR transcripts and provided metadata, whereas the Linking subtask required both audio and video features combnination. As Search sub-task had a predefined ground truth, a range of metrics was used for the evaluation of results [2]: Mean reciprocal rank (MRR) assesses the ranking of the relevant units; mean Generalized Average Precision (mGAP) awards runs that not only find the relevant items earlier in the list, but also are closer to the jump-in point of the relevant content; Mean Average Segment Precision (MASP) takes into account the ranking of the results and the length of both relevant and irrelevant segments that need to be listened to before reaching the relevant item. The result of Linking sub-task created new links between the videos within the collection that were assessed using MTurk platform only after participants submissions, and further standard Information Retrieval metric, mean Average Precision (mAP), was calculated.

4. SUMMARY

In this work, we first explained the need for tagging of social media and then we present a comprehensive multimedia dataset containing SPUG content, including the audiovisual content, user-contributed metadata, automatic speech recognition transcripts, automatic shot boundary files, and social information. The data set contains 14,838 episodes taken from `blip.tv` having a total runtime of ca. 3,288 hours and its size is about 862 GB. The size is divided between the single resources as follows: video files (ca. 764 GB), speech

transcripts (ca. 88 GB), shot boundary and key frame files (ca. 8 GB), social data (ca. 1.5 GB), and blip.tv's metadata (ca. 15 MB). This Blip10000 dataset was initially used in the MediaEval 2012 Tagging Task [14] and the Search and Hyperlinking Task [2]. These benchmark's participants showed promising results achieved using different dataset's resources. We hope research in the field of social media retrieval can be pushed forward by all resources offered by this dataset. To the best of our knowledge, this is the first media data set for tagging and search purposes that offers both video content with rich metadata and social network information. The dataset is available for downloading[8] since it is licensed under Creative Commons, except for the ASR transcripts—those can be only provided after signing usage agreement forms[9].

5. ACKNOWLEDGEMENTS

The work presented in this paper was partially supported by SEALINCMedia, Quaero and projects funded by the European Commission under contracts FP7-216444 Petamedia, FP7-261743 VideoSense, CUbRIK FP7-287704.

6. REFERENCES

[1] J. Almeida, T. Salles, E. Martins, O. Penatti, R. da S. Torres, M. Gonçalves, and J. Almeida. UNICAMP-UFMG at MediaEval 2012: Genre Tagging Task. In *Working Notes Proceedings of the MediaEval 2012 Workshop*. CEUR-WS.org, ISSN 1613-0073, October 4-5 2012.

[2] M. Eskevich, G. J. Jones, S. Chen, R. Aly, R. Ordelman, and M. A. Larson. Search and Hyperlinking Task at MediaEval 2012. In *Working Notes Proceedings of the MediaEval 2012 Workshop*. CEUR-WS.org, ISSN 1613-0073, October 4-5 2012.

[3] B. Ionescu, I. Mironică, K. Seyerlehner, P. Knees, J. Schlüter, M. Schedl, A. B. Horia Cucu, and P. Lambert. ARF @ MediaEval 2012: Multimodal Video Classification. In *Working Notes Proceedings of the MediaEval 2012 Workshop*. CEUR-WS.org, ISSN 1613-0073, October 4-5 2012.

[4] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui. Consumer video understanding: a benchmark database and an evaluation of human and machine performance. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 29:1–29:8, 2011.

[5] P. Kelm, S. Schmiedeke, and T. Sikora. Feature-based Video Key Frame Extraction for low Quality Video Sequences. In *10th Workshop on Image Analysis for Multimedia Interactive Services, 2009.*

[6] L. Lamel and J.-L. Gauvain. Speech processing for audio indexing. In B. Nordström and A. Ranta, editors, *Advances in Natural Language Processing*, volume 5221 of *Lecture Notes in Computer Science*, pages 4–15. Springer Berlin Heidelberg, 2008.

[7] 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 *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 51:1–51:8, New York, NY, USA, 2011. ACM.

[8] L. Y. Meng Wang and X.-S. Hua. MSRA-MM: Bridging Research and Industrial Societies for Multimedia Information Retrieval. In *TechReport: MSR-TR-2009-30*, 2008.

[9] M. Naaman. Social multimedia: highlighting opportunities for search and mining of multimedia data in social media applications. *Multimedia Tools Appl.*, 56(1):9–34, Jan. 2012.

[10] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. F. Smeaton, and G. Quéenot. TRECVID 2012 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proceedings of TRECVID 2012*. NIST, USA, 2012.

[11] K. K. Reddy and M. Shah. Recognizing 50 Human Action Categories of Web Videos. In *Machine Vision and Applications Journal (MVAP)*, 2012.

[12] A. Rousseau, F. Bougares, P. Deléglise, H. Schwenk, and Y. Estève. LIUM's systems for the IWSLT 2011 Speech Translation Tasks. In *International Workshop on Spoken Language Translation*, San Francisco (USA), 8-9 Sept 2011.

[13] S. Schmiedeke, P. Kelm, and T. Sikora. TUB @ MediaEval 2012 Tagging Task: Feature Selection Methods for Bag-of-(visual)-Words Approaches. In *Working Notes Proceedings of the MediaEval 2012 Workshop*.

[14] S. Schmiedeke, C. Kofler, and I. Ferrané. Overview of the MediaEval 2012 Tagging Task. In *Working Notes Proceedings of the MediaEval 2012 Workshop*. CEUR-WS.org, ISSN 1613-0073, October 4-5 2012.

[15] T. Semela, M. Tapaswi, H. K. Ekenel, and R. Stiefelhagen. KIT at MediaEval 2012 - Content-based Genre Classification with Visual Cues. In *Working Notes Proceedings of the MediaEval 2012 Workshop*.

[16] Y. Shi, M. A. Larson, P. Wiggers, and C. M. Jonker. MediaEval 2012 Tagging Task: Prediction based on One Best List and Confusion Networks. In *Working Notes Proceedings of the MediaEval 2012 Workshop*. CEUR-WS.org, ISSN 1613-0073, October 4-5 2012.

[17] X. Wu, A. G. Hauptmann, and C.-W. Ngo. Practical elimination of near-duplicates from web video search. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 218–227, New York, NY, USA, 2007. ACM.

[18] P. Xu, Y. Shi, and M. A. Larson. TUD at MediaEval 2012 genre tagging task: Multi-modality video categorization with one-vs-all classifiers. In *Working Notes Proceedings of the MediaEval 2012 Workshop*. CEUR-WS.org, ISSN 1613-0073, October 4-5 2012.

[19] S. Zanetti, L. Zelnik-manor, and P. Perona. A walk through the web's video clips. In *In: IEEE Workshop on Internet Vision, associated with CVPR*, 2008.

[8] `http://www.cngl.ie/MediaEval/Blip10000/Blip10000_downloadLinks.txt`

[9] `http://www.multimediaeval.org/mediaeval2012/tagging2012`