GeoUGV: User-Generated Mobile Video Dataset with Fine Granularity Spatial Metadata

Ying Lu† Hien To† Abdullah Alfarrarjeh† Seon Ho Kim† Yifang Yin‡
Roger Zimmermann‡ Cyrus Shahabi†

†Integrated Media Systems Center, University of Southern California, Los Angeles, CA 90089
‡ School of Computing, National University of Singapore, Singapore 117417
†{ylu720, hto, alfarrar, seonkim, shahabi}@usc.edu
‡{yifang, rogerz}@comp.nus.edu.sg

ABSTRACT

When analyzing and processing videos, it has become increasingly important in many applications to also consider contextual information, in addition to the content. With the ubiquity of sensor-rich smartphones, acquiring a continuous stream of geo-spatial metadata that includes the location and orientation of a camera together with the video frames has become practical. However, no such detailed dataset is publicly available. In this paper we present an extensive geo-tagged video dataset named GeoUGV that has been collected as part of the MediaQ [3] and GeoVid [1] projects. The key features of the dataset are that each video file is accompanied by a metadata sequence of geo-tags consisting of *GPS locations*, *compass directions*, and spatial keywords at *fine-grained* intervals. The GeoUGV dataset has been collected by volunteer users and its statistics can be summarized as follows: 2,397 videos containing 208,976 video frames that are geo-tagged, collected by 289 users in more than 20 cities across the world over a period of 10 years (2007–2016). We hope that this dataset will be useful for researchers, scientists and practitioners alike in their work.

CCS Concepts

•Information systems → Mobile information processing systems; Multimedia databases; Geographic information systems;

Keywords

Multimedia; Dataset; Mobile Video; Metadata; Geo-Tagging

1. INTRODUCTION

Due to the ubiquitous availability of smartphones, a number of trends have recently emerged with respect to mobile videos. First, we are experiencing unprecedented growth in the amount of mobile video content that is being collected with smartphones. Creating, searching, sharing and viewing videos are immensely popular activities with mobile users. This is facilitated by the ease with which it is possible to record and play back videos on mobile devices. Second, each smartphone includes a plethora of different built-in sensors such as cameras, a global position system (GPS) receiver, and a digital compass. This facilitates the modeling of video content through its geo-spatial properties at the fine granular level of a frame, a concept referred to as *Field-Of-View* (FOV) [6]. The FOV model has been shown to be very useful for various media applications such as online mobile media management systems, for example MediaQ [3, 12] and GeoVid [1, 5]. The GeoUGV dataset (http://mediaq.usc.edu/dataset, http://geovid.org/dataset) has been collected with their respective mobile apps [3, 1].

Offering a dataset of user-generated, geo-tagged videos can help researchers with various undertakings. (1) In advanced spatiotemporal video search, with the attached geo-metadata in GeoUGV, it is possible to convert the challenging problem of user-generated video indexing and querying to the problem of indexing and querying FOV spatial objects. Large-scale video data management using spatial indexing and querying of FOVs is a challenging problem, especially to maximally harness the geographical properties of FOVs. To attack this challenge, several indexes such as a grid-based index [18] and the OR-tree [16, 17] have been proposed. (2) Correlating video content with its spatial information adds a different perspective to video analytics (e.g., the acquisition of user behavior [24]). (3) 3D model reconstruction or creating panoramas at specific locations from user-generated videos can provide updated and "fresh" immersive user experiences. Spatial information can be utilized for effectively filtering irrelevant video frames [13, 25]. (4) The geo-metadata associated with videos can facilitate the selection of the most relevant videos/images for down-stream computer vision applications. For example, a persistent tracking application was studied with GIFT [7] (Geospatial Image Filtering Tool) to select the key video frames to significantly reduce the communication and processing costs in various computer vision applications. (5) Videos collected at different time for the same location enable the study of before and after situations for the location, which can be useful in applications such as disaster data analysis [23].

To the best of our knowledge, no existing public dataset contains user-generated, finely geo-tagged videos. Google Street View [2, 28] provides a set of images with locations

MMSys'16, May 10-13, 2016, Klagenfurt, Austria

DOI: http://dx.doi.org/10.1145/2910017.2910617

and compass directions captured using professional-grade equipment. However, Google data is collected by cars driving through every street once, which treats every location "equally" without reflecting the "popularity" of a place. However, in our user-generated dataset, more videos are collected in popular locations. Flickr provides an image dataset with locations. However, it does not provide camera viewing directions. The Stanford mobile video dataset [8] provides user-generated videos without spatial information. Arslan Ay et al. [4] have presented a synthesized dataset using a random walk model and it includes only metadata without the actual video content. On the other hand, GeoUGV is novel for the following two reasons: (1) it fused spatial metadata (such as GPS information and compass direction) with a set of sampled frames in the recorded video files and (2) the data were crowdsourced (also referred to as user-generated videos or UGVs) with a wide variety of mobile devices, i.e., it reflects a very heterogeneous set of hardware and software.

The remaining parts of this paper are organized as follows. Section 2 explains the data collection mechanism and describes the dataset. Section 3 provides statistics about our dataset followed by examples of the use of GeoUGV in Section 4. Finally, we conclude in Section 5.

2. DATASET

2.1 Spatial Model for Geo-Tagged Videos

We represent each video as a sequence of video frames, and the visible scene of each video frame is modeled as a Field of View (FOV) [6]. FOV is a comprehensive model to describe the scene a camera captures, which includes the camera view direction and camera location providing rich information to answer complex spatial video queries. As shown in Figure 1, each FOV f is in the form of $<p, \theta, R, \alpha>$, where p is the camera position consisting of the latitude and longitude coordinates read from the GPS sensor in a mobile device, θ is the angle of the camera viewing direction $\overrightarrow{d}$ with respect to the North obtained from the digital compass sensor, R is the maximum visible distance within which an object can be recognized, and α denotes the visible angle obtained from the camera lens property at the current zoom level. FOVs in two dimensions considering only camera azimuths are circular sector shaped while 3-dimensional FOVs are in cone shaped considering other two camera rotation types: pitch and roll. Note that the FOVs provided in GeoUGV are in two dimensions. We define $\mathcal{F}$ as the video frame set $\{f|\forall f \in v, \forall v \in \mathcal{V}\}$ in the set of all the videos, $\mathcal{V}$.

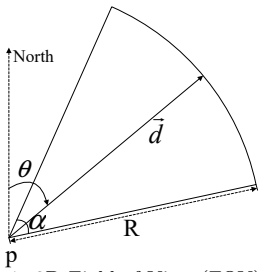


Figure 1: 2D Field-of-View (FOV) model.

2.2 Data Collection

We have collected the geo-tagged mobile video data from two mobile platforms, GeoVid [5, 1] and MediaQ [12, 3]. GeoVid is a system for collecting, indexing, searching geo-tagged videos. GeoVid was extended into MediaQ system which collects more metadata of video contents and additionally exploits the idea of spatial crowdsourcing termed GeoCrowd [11] to collect on-demand media content on behalf of users. GeoVid and MediaQ mobile apps (iOS or Android) are developed to collect videos and capture geo-metadata along with the videos. They share some common geospatial metadata, such as location and direction, in their own mobile videos collection.

While recording a video on a mobile device, various sensors (e.g., GPS, compass) are used to collect geospatial metadata (e.g., locations, camera viewing directions). Each time-stamped sensor data is recorded whenever its change is reported during the video recording. Note that we do not collect sensor readings if their values do not change over time. This *update-only* policy reduces the sampling rate significantly. Since the compass sensor generates readings very frequently, we limit its acquisition frequency (e.g., no more than 5 per second) to avoid unnecessary redundancy without losing accuracy. GPS sensor data is sampled approximately once per second. After recording a video, we generate the geo-metadata file by synchronizing both the sampled GPS and compass sensor data. Specifically, for each GPS record, we combine the compass reading with the closest timestamp. Then, we enhance the accuracy of the location metadata using a post-processing filtering step immediately after generating the metadata. We applied the data correction algorithm with Kalman filtering and weighted linear least square regression [26] to manage potential big variance in GPS signals.

2.3 Dataset Description

The GeoUGV dataset consists of two sets, videos and their geospatial metadata. The metadata is stored in two files; the first file (i.e., *VideoMetadata.txt*) contains the video file and the recording mobile device information for each video, and the second file (i.e., *FOVMetadata.txt*) includes the geo-information about the sampled geo-frames (i.e., FOVs) of videos. Both metadata files are in a tab space separated value format.

2.3.1 Video Metadata

Each record in the *VideoMetadata.txt* file contains the information of a video. Each record is composed of seven fields:

- $VideoFileName$: the physical video file name which is considered as the identifier of the video record.
- $DeviceID$: an identifier of the smartphone used for recording the video. In our system, we used the subscriber identity module (SIM) identifier to generate a unique identifier for a smartphone. For privacy reasons, we reported anonymized device IDs to protect user identity.
- $DeviceModel$: the model of the smartphone. Example values of this field include *Galaxy Nexus* , *GT-I9190*, and *Nexus 5*.
- $IsVideoContentUploaded$: a flag indicating whether the actual video file exists in GeoUGV. Because the

size of metadata is small, our app automatically sends it to the server immediately after a video recording is done and the user can choose to upload the video later whenever a good network bandwidth is available. In GeoUGV, 19.73% of the total videos have only metadata without video files. The value of this field is either *1* or *0*; *1* means both the actual video and its metadata exist and *0* means only the video metadata is available.

- $VideoResolution$: the resolution of recorded video in the number of pixels (e.g., 1920×1080).
- $VideoLength$: the temporal length of the video file in seconds.
- $FrameCount$: the total number of frames included in the video file.

The values of the first five fields are captured during video recording while the remaining fields are computed for records with both metadata and actual video files after processing the video files. The data type of all the fields is string (i.e., a sequence of characters) except the fields $VideoLength$, $IsVideoContentUploaded$, and $FrameCount$ which are numeric values. There are 2397 records in the *VideoMetadata.txt* file, and the size of the file is 210 kilobytes.

2.3.2 Spatial Video Metadata

In the metadata file (i.e., *FOVMetadata.txt*), we provide a fine granularity geospatial metadata about the collected videos, consisting of a set of frame records (i.e., FOVs). Each record represents an FOV comprised of nine fields:

- $VideoFileName$: the video file name that contains the FOV.
- $FOVNumber$: the sequential number of an FOV in the video named $VideoFileName$. An FOV can be uniquely identified with the tuple ($VideoFileName$, $FOVNumber$).
- *Latitude*: the latitude of the camera location, based on the GPS coordinates, when the frame was recorded. The value of *Latitude* is in the range [−90°, 90°] in double precision (i.e., rounded off up to 15 decimal places).
- *Longitude*: the longitude of the camera location. The value of *Longitude* is within the range [−180°, 180°] in double precision.
- θ: the azimuth angle of the camera viewing direction. It expresses the angular distance from the north, as shown in Figure 1. The angle θ is a double precision value within the range [0°, 360°].
- R: is the maximum visible distance of the FOV within which an object can be recognized. In GeoUGV, the value of R is a predefined constant (e.g., 0.1 km).
- α: the visible angle of the FOV obtained from the camera lens property at the zoom level when the frame is recorded. The value of α is also set as a predefined constant (e.g., 51°).
- $Timestamp$: the number of elapsed milliseconds from 1970-01-01 00:00:00.

Total number of videos with geo-metadata	2,397
Total number of videos with both geo-metadata and contents	1,924
Total length of videos with contents (hour)	38.54
Average length per video with content (sec)	72.14
Percentage of videos which have keywords	22.78%
Average camera moving speed (km/h)	4.5
Average camera rotation speed (degrees/sec)	10
Total number of users	289
Average number of videos by each user	8.29
Total number of FOVs	208,976
Total FOV number of video with contents	142,687
Average number of FOV per second	1.03
Average number FOV per video	74.16

Table 1: Overview of the GeoUGV dataset.

- $Keywords$: a set of keywords that are associated with the FOV. The keywords are separated by a semicolon. (e.g., school; new high street; spring street). Our apps provide two types of keywords: automatic and manual. The automatic keywords are extracted from OpenStreetView[1] based on the FOV covering area [20]. The manual keywords are provided by users when uploading a video. The manual keywords for a video are populated to all FOVs belonging to it. If an FOV is associated with multiple keywords, then the keywords are separated with commas. If there is no keyword, the field is NULL.

There are 208,976 records in the *FOVMetadata.txt* file, and the size of the file is 26.4 megabytes.

3. MOBILE VIDEO STATISTICS

In this section, we present the characteristics and statistics of our GeoUGV dataset.

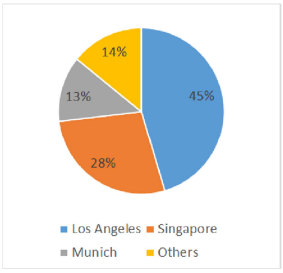
Table 1 shows the overall statistics of GeoUGV collected in the past 10 years ($2007 - 2016$). There are 2397 videos in total, of which 1924 videos have both video contents and their geospatial metadata. The total length of the 1924 videos is 38.54 hours with 58.18 gigabytes in size, and the average video length is 72.14 seconds. As we discussed in Sec. 2.3.2, videos can be attached with manually typed keywords and/or automatically tagged keywords based on their FOV coverage. In our dataset, 22.78% of the videos have keywords. Most of the videos were recorded by users casually in a walk mode. The camera moving speed is 4.5 km/h on average, and the camera rotation speed is 10 degrees/sec (i.e., the orientation θ changing speed). In addition, the GeoUGV dataset were collected by 289 users, and each user collected 8.29 videos on average. Moreover, there were a total of 208,976 FOVs available in GeoUGV, among which 142,687 FOVs are associated with video contents as well. Therefore, the average sampling rate is 1.03 FOVs per second, and each video is attached with 74.16 FOVs on average.

To illustrate the places where the videos were taken, Figure 2 displays the location distribution of the collected videos on Google Maps. As can be seen, the videos were collected from all over the world. Specifically, as shown in Figure 3(a), most of the videos were recorded in Los Angeles (45%), Singapore (28%), Munich (13%), and the rest 14% videos were from 18 other cities. We also plotted the chronological distribution of the number of videos collected ($2007 - 2016$) in Figure 3(b). Most of the data were collected during the recent five years ($2011 - 2015$). There were 289 users who participated in this dataset collection, and the number of

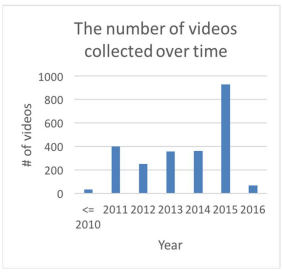
[1] http://openstreetview.org/

videos collected by each user is summarized in Figure 3(c). Around one third of the users recorded only one video while 18% of the users collected more than 10 videos. We also present the numbers of videos collected by mobile devices with different OS (e.g., Android, iOS) in Figure 3(d). Note that the OS types of 62% of the videos are unknown because the old version of our app did not collect this information. Figure 4 shows the distribution of video lengths. It is worth noting that more than 40% of the collected videos were less than 30 seconds long, and around 4% of the videos were more than 300 seconds. Furthermore, as shown in Figure 5, the resolutions of the majority of the videos were 640×480 (36%) and 720×480 (50%), which have been the standard resolutions of digital videos on smartphones.

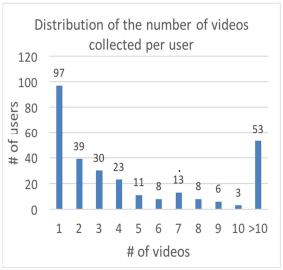
Figure 2: Locations of the videos.



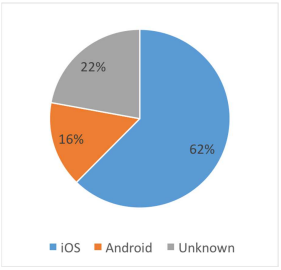
(a) Per city



(b) Per year



(c) Per user



(d) Per OS type

Figure 3: The distributions of the number of videos.

	Places	Video#
$p1$	Near RTH Building @ USC, LA, USA	13
$p2$	37th Street on USC Campus @ LA, USA	10
$p3$	Chinesischer Turm @ Munich, Germany	8
$p4$	Tommy Trojan @ USC, LA, USA	7
$p5$	Near Leavey Library @ USC, LA, USA	6

Table 2: Top 5 video dense locations

We analyzed the geospatial distribution of the collected videos. In some popular places, there were many videos as

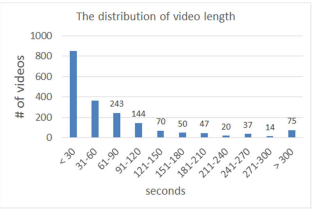


Figure 4: Video length distribution

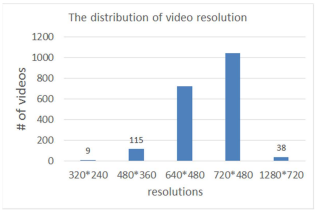


Figure 5: Video resolution distribution

	Places	FOV#
$p6$	Merlion Park @ Singapore	293
$p7$	Esplanade Theatres @ Singapore	234
$p8$	Singapore Cruise Center @ Singapore	156
$p9$	The Float @ Marina Bay, Singapore	138
$p10$	Bedok Reservoir Park @ Singapore	101

Table 3: Top 5 FOV dense locations

we can easily expect. Theses places include the geo-tagged video dataset which were used in our previous study of automatic generation of panoramic images [13] and 3D model reconstruction [25] from the collected geo-tagged videos. To find dense locations (i.e., places with many videos/FOVs), we calculate the numbers of videos and FOVs that were located within a certain "place" (i.e., a 100 meters $\times$ 100 meters cell) in the GeoUGV dataset. Note that the videos we consider in the calculation contain both video contents and geo-metadata. Specifically, for each video, we use the camera location (*Latitude*, *Longitude*) of its first FOV as the location of the video. We refer the camera location of FOV as the FOV location. After the calculation, we find there were 14 places with more than 5 videos, and the top 5 video densest places are listed in Table 2. Furthermore, there were 62 places with more than 20 FOVs, and the top 5 FOV densest places are listed in Table 3. In Table 2, four of the places are within the USC campus in Los Angeles and the other one is in Munich. In Table 3, all the places are in Singapore. Their top-5 densest places are different mainly because the videos collected in Singapore were longer. Note that these densely populated places are of special interest since there can be enough amount of visual data to be used for panorama generation, motion structure, 3D model reconstruction, etc., which might be a good source of visual data for researchers in computer vision. For example, the videos collected in $p2$, a quiet street on the USC campus, were used in automatically generating route panoramic images in the authors' previous work [13]. Similarly, the videos collected in $p3$, $p5$, $p6$, $p7$ and $p9$ where are open areas were used in generating point panorama images [13]. The videos recorded in $p4$ and $p6$, where some recognizable figures (e.g., famous statures like Tommy Trojan at USC, Merlion at Merlion Park in Singapore) are located, were used in automatic 3D model reconstruction [25].

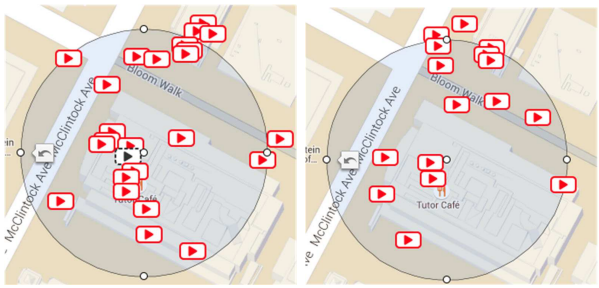
4. EXAMPLES OF USE

The GeoUGV dataset can be used for a variety of purposes. This section provides some example use cases of the spatial metadata by the authors.

4.1 Advanced Spatiotemporal Video Search

It is very challenging to index and search videos and pictures at a large scale. Traditional techniques, such as anno-

tating videos with keywords and content-based retrieval, are often unsatisfactory due to the lack of appropriate keywords and the inaccuracy of search results, respectively. While manually annotated videos can be efficiently retrieved, annotating large video collections is laborious and time-consuming. On the other hand, content-based video retrieval is challenging, computationally expensive and the result is often associated with uncertainty. With the availability of sensor-rich metadata (e.g., location, direction), it is efficient to search video data at the high semantic level preferred by humans [12]. For example, region queries (i.e., rectangular query or circle query) can easily retrieve all FOVs that overlap with a given region using spatial indexing technologies. A directional query searches all video segments whose FOV direction angles are within a range of an allowable error margin to a user-specified input direction angle. Figure 6 shows different results of two directional range queries whose spatial regions are the same.



(a) North (b) South

Figure 6: Examples of directional query on MediaQ. The red icons represent video segments with a specific compass direction.

4.2 Video Data Analytics

Using geospatial metadata can facilitate video data analysis. For example, one can detect interesting events (or hotspots) from a large number of videos [24]. Figure 7 shows FOV coverage of 315 videos collected during two years 2014 and 2015 at the University of Southern California. The figure shows that several hotspots are automatically recognized. There are two types of hotspots, points of interest where many people visited frequently (i.e., RTH and PHE buildings and Tommy Trojan) and actual events happened at the campus where many videos were collected during a short time period (e.g., two-day LA book festival). Detecting hotspots is also helpful in efficient disaster management [23]. Particularly, the video coverage map computed with metadata would help decision makers to understand the disaster situation quickly and provide timely actions accordingly. For example, on-site volunteers can be sent to the hotspots to check the situation or off-site analysis can collect more data in sparsely covered areas using spatial crowdsourcing.

Another application regarding metadata analytics is that one may quickly detect informative video segments (i.e., a sequence of FOVs with more useful visual information) from a large collection of videos. The study [24] shows that significant video segments can be identified by searching for specific camera shooting patterns defined in filming, such as panning, zooming, tracking and arching. One may define a significant segment using a specific pattern. For example, if a user stays at a location and records a video for a

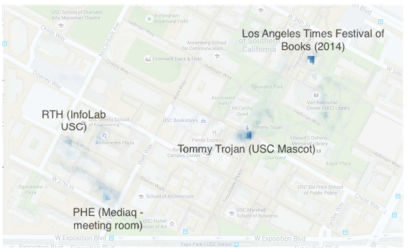


Figure 7: A heatmap of FOV coverage of 315 videos are overlaid on top of Google Maps. This shows that it is possible to automatically detect hotspots from geo-tagged mobile video metadata.

while pointing his camera toward a specific direction, it is most likely that there is an interesting event in front of him. Therefore, by searching for a specific pattern (e.g., a period of at least 20 seconds while movement is minimal) in camera movements, we might be able to efficiently find a set of informative video segments. Figure 8 gave a event example we identified where someone stops and records a group of people singing on a stage. The pattern can be relaxed to search for panoramic scenes (i.e., the camera location does not change while changing the shooting direction) [13].

Furthermore, by analyzing these geo-tagged mobile videos taken by public users with moving trajectories and camera shooting directions, we can utilize the GeoUGV dataset for trip recommendation [15]. For example, a recent study [14] mines the FOV metadata of photos to extract the frequent travel patterns of users and returns high-quality personalized itineraries.

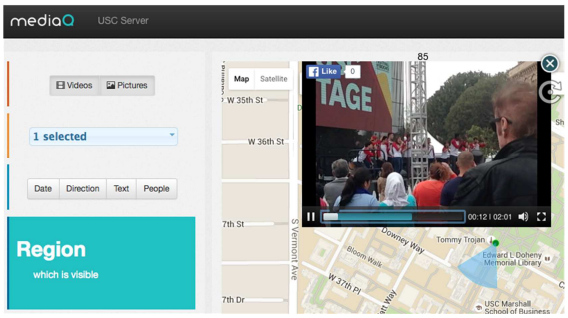


Figure 8: An actual event automatically found from a large collection of geo-tagged mobile videos.

4.3 Geospatial Filtering for Computer Vision Applications

In computer vision field, researchers mainly focus on the analysis of a given set of multimedia documents without considering what would be the most effective input for the analysis. Therefore, when an unprecedented number of videos and images are currently being collected, how to efficiently select the relevant input videos and images for computer vision applications becomes a challenging problem. The geo-metadata associated with videos in GeoUGV can facilitate the selection of the most relevant video segments or images for down-stream computer vision applications (e.g., object tracking, feature extraction, people counting). For example, a persistent tracking application has been studied with GIFT [7] (Geospatial Image Filtering Tool) to select the key video frames, which significantly reduced the communication and processing costs.

4.4 Content-based video retrieval

Even though video search has been widely studied in the research community, identifying an object in a video database is still a challenging problem. Searching videos has been studied in different paradigms: content-based, keyword-based, semantic-based, and spatial-based. In content-based search, low-level features (e.g., color histograms, Gabor texture features, and motion vectors) are usually extracted for visual similarity search [21]. In addition to user-generated annotations [22, 9], various context information has also been utilized to probe the semantics of video content [19, 10]. As one of such important context clues, spatial metadata enables efficient video search [27, 16, 17, 18]. Our dataset consisting of geo-tagged videos with automatically annotated spatial keywords, improves video retrieval when leveraging fusion paradigms in video search.

5. CONCLUSION & FUTURE WORK

In this paper, we have presented a new dataset of user-generated mobile videos. The key feature of this dataset is that each video file is accompanied with a metadata sequence of geo-tags such as GPS *locations*, *compass directions*, and spatial keywords at *fine-grained* intervals. To the best of our knowledge, no existing public dataset contains user-generated, finely geo-tagged videos. With the additional spatial information, this dataset can be used in advancing spatiotemporal video search, accelerating video content analysis technologies, video mining and video retrieval and ranking. Future extensions of this dataset will mainly target the following directions: 1) providing more precise camera setting properties, e.g., zoom level, camera lens; 2) incorporating with social information, e.g., the number of users that view, like and share the videos, for social media applications.

6. ACKNOWLEDGMENTS

This research has been funded in part by NSF grants IIS-1115153, IIS-1320149, and CNS-1461963, the USC Integrated Media Systems Center (IMSC), and unrestricted cash gifts from Google, Northrop Grumman, Microsoft, and Oracle. This research has also been supported by the Singapore National Research Foundation under its International Research Centre Singapore Funding Initiative and administered by the IDM Programme Office through the Centre of Social Media Innovations for Communities (COSMIC). In addition, we would like to thank Tobias Emrich and Matthias Renz for contributing the data in Munich.

7. REFERENCES

[1] GeoVid Project. http://geovid.org/.

[2] Google Street View Dataset. http://crcv.ucf.edu/data/GMCP_Geolocalization/#Dataset.

[3] MediaQ Project. http://mediaq1.cloudapp.net/home/.

[4] S. Arslan Ay, S. H. Kim, and R. Zimmermann. Generating Synthetic Meta-data for Georeferenced Video Management. In *ACM SIGSPATIAL GIS*, pages 280–289, 2010.

[5] S. Arslan Ay, L. Zhang, S. H. Kim, M. He, and R. Zimmermann. GRVS: A Georeferenced Video Search Engine. In *17th ACM International Conference on Multimedia*, pages 977–978, 2009.

[6] S. Arslan Ay, R. Zimmermann, and S. H. Kim. Viewable Scene Modeling for Geospatial Video Search. In *16th ACM International Conference on Multimedia*, pages 309–318, 2008.

[7] Y. Cai, Y. Lu, S. H. Kim, L. Nocera, and C. Shahabi. Gift: A Geospatial Image and Video Filtering Tool for Computer Vision Applications with Geo-tagged Mobile Videos. In *Multimedia and Expo Workshops*, pages 1–6, 2015.

[8] V. R. Chandrasekhar, D. M. Chen, S. S. Tsai, N.-M. Cheung, H. Chen, G. Takacs, Y. Reznik, R. Vedantham, R. Grzeszczuk, J. Bach, et al. The Stanford Mobile Visual Search Data Set. In *2nd ACM Conference on Multimedia Systems*, pages 117–122, 2011.

[9] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu. Visual-textual Joint Relevance Learning for Tag-based Social Image Search. *IEEE Transactions on Image Processing*, 22(1):363–376, 2013.

[10] L. Jiang, S.-I. Yu, D. Meng, T. Mitamura, and A. G. Hauptmann. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In *5th ACM International Conference on Multimedia Retrieval*, pages 27–34, 2015.

[11] L. Kazemi and C. Shahabi. GeoCrowd: Enabling Query Answering with Spatial Crowdsourcing. In *ACM SIGSPATIAL GIS*, pages 189–198, 2012.

[12] S. H. Kim, Y. Lu, G. Constantinou, C. Shahabi, G. Wang, and R. Zimmermann. MediaQ: Mobile Multimedia Management System. In *5th ACM Conference on Multimedia Systems*, pages 224–235, 2014.

[13] S. H. Kim, Y. Lu, J. Shi, A. Alfarrarjeh, C. Shahabi, G. Wang, and R. Zimmermann. Key Frame Selection Algorithms for Automatic Generation of Panoramic Images from Crowdsourced Geo-tagged Videos. In *Web and Wireless Geographical Information Systems*, pages 67–84. Springer, 2014.

[14] C.-C. Lin, Y. Zhang, Y.-L. Hsueh, and R. Zimmermann. A Personalized Trip Recommendation System Based on Field of Views. In *5th International Conference on Engineering and Applied Sciences*, 2015.

[15] Y. Lu and C. Shahabi. An Arc Orienteering Algorithm to Find the Most Scenic Path on a Large-scale Road Network. In *23rd ACM SIGSPATIAL GIS*, pages 46:1–46:10, 2015.

[16] Y. Lu, C. Shahabi, and S. H. Kim. An Efficient Index Structure for Large-scale Geo-tagged Video Databases. In *ACM SIGSPATIAL GIS*, pages 465–468, 2014.

[17] Y. Lu, C. Shahabi, and S. H. Kim. Efficient Indexing and Retrieval of Large-scale Geo-tagged Video Databases. *GeoInformatica*, pages 1–29, Apirl 2016.

[18] H. Ma, S. Arslan Ay, R. Zimmermann, and S. H. Kim. Large-scale Geo-tagged Video Indexing and Queries. *GeoInformatica*, 2013.

[19] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative Multi-label Video Annotation. In *15th ACM International Conference on Multimedia*, pages 17–26, 2007.

[20] Z. Shen, S. Arslan Ay, S. H. Kim, and R. Zimmermann. Automatic Tag Generation and Ranking for Sensor-rich Outdoor Videos. In *19th ACM International Conference on Multimedia*, pages 93–102, 2011.

[21] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *9th IEEE International Conference on Computer Vision*, pages 1470–1477, 2003.

[22] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua. Bayesian video search reranking. In *16th ACM International Conference on Multimedia*, pages 131–140, 2008.

[23] H. To, S. H. Kim, and C. Shahabi. Effectively Crowdsourcing the Acquisition and Analysis of Visual Data for Disaster Response. In *IEEE Big Data*, pages 697–706, 2015.

[24] H. To, H. Park, S. H. Kim, and C. Shahabi. Incorporating Geo-Tagged Mobile Videos Into Context-Aware Augmented Reality Applications. In *The 2nd IEEE International Conference on Multimedia Big Data*, 2016.

[25] G. Wang, Y. Lu, L. Zhang, A. Alfarrarjeh, R. Zimmermann, S. H. Kim, and C. Shahabi. Active Key Frame Selection for 3D Model Reconstruction from Crowdsourced Geo-tagged Videos. In *International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2014.

[26] G. Wang, B. Seo, and R. Zimmermann. Automatic Positioning Data Correction for Sensor-annotated Mobile Videos. In *ACM SIGSPATIAL GIS*, pages 470–473, 2012.

[27] Y. Yin, Y. Yu, and R. Zimmermann. On Generating Content-Oriented Geo Features for Sensor-Rich Outdoor Video Search. *IEEE Transactions on Multimedia*, 17(10):1760–1772, 2015.

[28] A. R. Zamir and M. Shah. Image Geo-Localization Based on MultipleNearest Neighbor Feature Matching UsingGeneralized Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1546–1558, 2014.