ACTIVE KEY FRAME SELECTION FOR 3D MODEL RECONSTRUCTION FROM CROWDSOURCED GEO-TAGGED VIDEOS

Guanfeng Wang[†]*, Ying Lu*[‡]*, Luming Zhang*[§]*, Abdullah Alfarrarjeh*[‡]
Roger Zimmermann[†]*, Seon Ho Kim*[‡]*, Cyrus Shahabi*[‡]

[†§]School of Computing, National University of Singapore, Singapore 117417
[†]{wanggf,rogerz}@comp.nus.edu.sg [§]zglumg@gmail.com
[‡]Integrated Media Systems Center, University of Southern California, Los Angeles, CA 90089
[‡]{seonkim, ylu720, alfarrar, shahabi}@usc.edu

ABSTRACT

Automatic reconstruction of 3D models is attracting increasing attention in the multimedia community. Scene recovery from video sequences requires a selection of representative video frames. Most prior work adopted content-based techniques to automate key frame extraction. However, these methods take no frame geo-information into consideration and are still compute-intensive. Here we propose a new approach for key frame selection based on the geographic properties of videos. Currently, an increasing number of user-generated videos (UGVs) are collected – a trend that is driven by the ubiquitous availability of smartphones. Additionally, it has become easy to continuously acquire and fuse various sensor data (*e.g.*, geo-spatial metadata) with video to create *geo-tagged* mobile videos. Our novel technique utilizes these underlying geo-metadata to select the most representative frames. Specifically, a key frame subset with minimal spatial coverage gain difference is extracted by incorporating a manifold structure into reproducing a kernel Hilbert space to analyze the spatial relationship among the frames. Our experimental results illustrate that the execution time of the 3D reconstruction is shortened while the model quality is preserved.

Index Terms— Key frame selection, 3D reconstruction, manifold adaptive kernel space, geo-tagged mobile video.

1. INTRODUCTION

Recently there has been significant progress in techniques that focus on recovering 3D scene geometry from multiple 2D images. Traditionally, the images used for such purposes are carefully and specifically recorded to show the candidate 3D object from different viewing angles while avoiding too much spatial overlap. It is important to obtain a near-optimal number of images from distinct viewing angles because too few images will result in visual "holes" in the reconstructed object, while too many images will unnecessarily increase the computational load and execution time and in some cases introduce artifacts.

Recently, many scenes (especially outdoors) are being captured from multiple viewpoints through user-generated videos (UGV). The ubiquituous availabilty of smartphones with video recording capabilities has made this trend possible. In this work we explore the feasibility of using a set of UGVs to reconstruct 3D objects within an area. Such a method introduces the following challenges. First, videos are recorded at 25 or 30 frames per second and successive frames are very similar. Hence not all video frames should be used – rather, a set of *key frames* needs to be extracted that provide optimally sparse coverage of the candidate object. Second, the camera position and visual trajectory of UGVs are determined by the actions of an individual user. Such videos are not usually captured with 3D reconstruction in mind. To overcome this issue we leverage another technological trend. Current smartphones contain sensors that can capture the geographic properties of the recorded scene, specifically the camera position (through GPS) and the viewing direction (via a digital compass). Such geo-spatial metadata can automatically be attached to videos at a fine-granular frame level and then utilized to select an effective set of key frames.

The main component in the proposed 3D object reconstruction framework is an active key frame selection algorithm. To efficiently determine an effective set of key frames, we leverage (a) the available crowdsourced UGVs in the region and (b) the frame-attached geo-spatial metadata. In effect, our approach enables the "re-purposing" of UGVs for 3D object reconstruction. Our experimental results demonstrate not only the computational feasibility of the proposed method but also the output quality of the generated 3D models.

2. RELATED WORK

3D model reconstruction from images [1, 2, 3] or videos [4, 5] has been of wide interest to the research community. Other researchers [6, 7, 8] have considered selecting key frames from

a video prior to initiating the reconstruction process. The existing selection techniques extract key frames from one video source, while we propose selection techniques from multiple crowdsourced UGVs. In the method from Ahmed *et al.* [6], the selection mechanism of key frames is based on (a) the number of frame-to-frame point correspondences obtained from a geometrical robust information criterion (GRIC) [9], and (b) the point-to-epipolar line cost for the frame-to-frame correspondence set. Other work [8] considers more factors to select key frames: the ratio of the number of point correspondences found to the total number of point features found, the homography error, and the spatial distribution of corresponding points over the frames. Seo *et al.* [7] use the ratio of the number of correspondences to the total number of features found. One of the common characteristics of the existing techniques is that they select key frames depending on different geometric models to score the correspondence of matching points between frames. All these methods focus on the frame content- or point cloud-level processing which are still compute-intensive, while our method instead focuses on UGV attached sensor data to choose the most representative key frames in geographic space.

Mordohai *et al.* [10] also used GPS data within a real time 3D reconstruction approach from videos that makes use of location information to place the reconstructed models in geo-registered coordinates on maps. However, their acquisition system needs to be fully customized and they simply select the candidate frames whose baseline between two consecutive frames exceeds a certain threshold for further 3D reconstruction. They also mention that the threshold varies depending on the different objects' scene depth. Instead, we employ GPS information and more sophisticated algorithms to select a set of geographically representative frames of the collected videos. To the best of our knowledge, there exists no prior method that leverages crowdsourced videos that are contextually enriched at a very fine-grained level and extracts key frames based on their geographic characteristics to reconstruct 3D models.

3. GEO-BASED ACTIVE KEY FRAME SELECTION

The geo-sensor data utilized in our approach is a series of contextual descriptions of mobile video content that reflects the geospatial properties of the captured scenes. We adopt the field-of-view (FOV, also called the *viewable scene*) model [11], which consists of three parameters: the camera location $\boldsymbol{p}$, the camera orientation θ, and the viewable angle α, $FOV \equiv \langle \boldsymbol{p}, \theta, \alpha \rangle$. To allow users to conveniently acquire geo-tagged videos, we leverage a custom recording app named *GeoVid*, publicly available for both Android and iOS [12, 13]. Note, each mobile device model may use different sampling frequencies for different sensors. Ideally we acquire one *FOV triplet* per frame. If that is not feasible and the granularity is coarser due to the device limits, we perform linear interpolation to generate triplets for each frame.

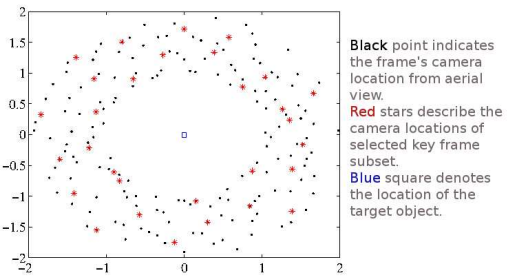


Fig. 1. Illustration of geo-based active key frame selection algorithm in 2D space.

With the FOV model, each frame corresponds to a camera geo-location and orientation. Here we focus on the frame's location and viewing direction in the geographic domain instead of the traditional pixel domain to extract the most representative frames for 3D reconstruction. Since some UGVs in the candidate area are not recording the target object, we first filter out all the frames that do not contain the target

$$\langle \boldsymbol{p}_i, \theta_i, \alpha_i \rangle : \|D(\boldsymbol{p}_i, \boldsymbol{p}_q) - \theta_i\| \leq \frac{\alpha}{2} \tag{1}$$

where $\langle \boldsymbol{p}_i, \theta_i, \alpha_i \rangle$ is the FOV triplet of the i^{th} frame, $\boldsymbol{p}_q$ is the geo-location of the target object, and D is a direction function that calculates the viewing direction, given two positions. As illustrated in Figure 1, the black points are the frames' camera locations from an aerial view. Without loss of generality, we assume that those frames record the object in the center (denoted with a blue square) after the filtering phase. The objective of our algorithm is to select a subset of frames (denoted with red stars), which maintain a minimal, but full, coverage of the target object in the geographic space. In other words, the information loss from any viewing angle towards the target object is minimized by our key frame selection method.

Let $\mathcal{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n\}$ be the set of all frame geo-location data points and $\mathcal{K} = \{\boldsymbol{k}_1, \boldsymbol{k}_2, \ldots, \boldsymbol{k}_t\} \subset \mathcal{P}$, termed *key location set* in this paper, be the set of the selected location points. Each location data point consists of a coordinate in the World Geodetic System 1984 [14]. Here we propose a coverage gain function in the geographic space, $g(\boldsymbol{p}) = \boldsymbol{w}^T \boldsymbol{p}$, to quantify the target object's viewing angle coverage. Suppose $l = g(\boldsymbol{p}) + \epsilon$ is a measurable label from the geographic coverage relation between $\boldsymbol{p}$ and the target object's location $\boldsymbol{p}_q$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is the measurement error. Thus, the maximum likelihood estimate of w can be obtained by

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \sum_{i=1}^{t} (\boldsymbol{w}^T \boldsymbol{k}_i - l_i)^2 \tag{2}$$

The key idea of our selection approach is to minimize the difference between the coverage gain based on all frame locations and the *key location set*. Specifically, the average ex-

pected square difference of the estimation function g needs to be minimized. We start by stating the problem formally.

Problem Statement. Given a set of frames $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$ and their corresponding geo-locations $\mathcal{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n\}$, find a key frame subset $\tilde{\mathcal{F}}$ whose corresponding geo-locations are $\mathcal{K} = \{\boldsymbol{k}_1, \boldsymbol{k}_2, \ldots, \boldsymbol{k}_t\} \subset \mathcal{P}$, and the average expected square coverage gain difference, $G_{diff} = \frac{1}{n}\sum_{i=1}^{n} E(l_i - \hat{l})^2$, is minimized, with $\hat{l} = \hat{\boldsymbol{w}}^T \boldsymbol{p}$.

We derive the expected square coverage gain difference as follows:

$$\begin{aligned} G_{diff} &= \frac{1}{n}\sum_{i=1}^{n} E(\hat{\boldsymbol{w}}^T \boldsymbol{p}_i - l_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{p}_i{}^T [E(\boldsymbol{w} - \hat{\boldsymbol{w}})(\boldsymbol{w} - \hat{\boldsymbol{w}})^T] \boldsymbol{p}_i \end{aligned} \tag{3}$$

By the Gauss-Markov theorem, the covariance matrix of $(\boldsymbol{w} - \hat{\boldsymbol{w}})$ is σ^2 times the inverted Hessian of $\sum_{i=1}^{t}(\boldsymbol{w}^T \boldsymbol{k}_i - l_i)^2$. So G_{diff} can be written as

$$G_{diff} = \sigma^2 + \sigma^2 \mathrm{Tr}(P^T (KK^T)^{-1} P)$$

where $P = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n)$ and $K = (\boldsymbol{k}_1, \ldots, \boldsymbol{k}_t)$. We find that the measurement l does not appear in the equation, so the average expected square coverage gain difference only depends on *key location set* $\mathcal{K}$. This mathematical structure and its semantic objective can be formulated as Transductive Experimental Design (TED), an active learning model from the machine learning community [15, 16, 17]. This problem is often referred to as experiment design in statistics [18] and such an optimization has been verified as being an NP-hard problem [19]. We employ a convex relaxation of the minimization problem recently proposed by Yu *et al.* [20]:

$$\min_{\beta, \boldsymbol{\alpha}_i \in \mathbb{R}^n} \sum_{i=1}^{n} \|\boldsymbol{p}_i - K^T \boldsymbol{\alpha}_i\|^2 + \sum_{j=1}^{n} \frac{\alpha_{i,j}^2}{\beta_j} + \gamma \|\boldsymbol{\beta}\|_1$$

where $\boldsymbol{\alpha}_i = (\alpha_{i,1}, \ldots, \alpha_{i,n})^T$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$ are the auxiliary variables to control the inclusion of examples into the *key location set*. This has been proved to be a convex problem and a global optimal solution is guaranteed. All candidates with $\beta_j = 0$ can be rejected, since the l_1-norm $\|\boldsymbol{\beta}\|$ enforces a sparse $\boldsymbol{\beta}$.

To achieve the best viewing angle coverage around the target object in geographic space, intuitively we need to take the geometric structure of the data points into consideration. Thus we adopt the Manifold Adaptive Kernel [21] which incorporates the manifold structure into the reproducing kernel Hilbert space (RKHS) to reflect the underlying geometry of the data. To model the structure, we also construct a nearest neighbor graph whose weight matrix element W_{ij} is 1 if two data points exist within each other's t nearest neighbors [22]. The graph Laplacian accordingly is defined as $L = W' - W$ where W' is given by $W'_{ii} = \sum_j W_{ij}$. Finally we obtain the manifold adaptive reproducing kernel as:

$$\mathcal{K}_{\mathcal{M}}(\boldsymbol{p}, \boldsymbol{k}) = \mathcal{K}(\boldsymbol{p}, \boldsymbol{k}) - \lambda \boldsymbol{s}_{\boldsymbol{p}}^T (I + LH)^{-1} L \boldsymbol{s}_{\boldsymbol{k}}$$

where $\boldsymbol{s}_{\boldsymbol{p}} = (\mathcal{K}(\boldsymbol{p}, \boldsymbol{p}_1), \ldots, \mathcal{K}(\boldsymbol{p}, \boldsymbol{p}_n))$, I is an identity matrix, λ is a constant controlling the smoothness of the functions and H is the kernel matrix in $\mathcal{H}$. $\mathcal{H}$ is a complete Hilbert space of functions $\mathcal{E} \to \mathbb{R}$, where $\mathcal{E}$ is a compact domain in a Euclidean space or a manifold [21]. Cai *et al.* have shown that this optimization problem can be solved by performing a convex TED in manifold adaptive kernel space [22]. We utilize their model by initializing $\alpha_{i,j} = 1$ and iteratively computing

$$\beta_j = \sqrt{\frac{\sum_{i=1}^{n} \alpha_{i,j}^2}{\gamma}}, \quad j = 1, \ldots, n,$$

$$\boldsymbol{\alpha}_{\boldsymbol{i}} = (diag(\beta)^{-1} + H)^{-1} \boldsymbol{u}_{\boldsymbol{i}}, \quad i = 1, \ldots, n,$$

until convergence, where $\boldsymbol{u}_{\boldsymbol{i}}$ is the i^{th} column vector of H and $H_{ij} = \mathcal{K}(\boldsymbol{p}_{\boldsymbol{i}}, \boldsymbol{p}_{\boldsymbol{j}})$. The data points afterwards can be ranked in a descending order with regard to β_j and then selecting the top t as *key location set* $\mathcal{K}$.

4. EXPERIMENTS

The main purpose of our key frame selection strategy is to maintain as much view coverage of the target object as possible with the minimally necessary number of frames. Therefore, we focus on two aspects in our experimental evaluation: (a) the geographic coverage gain difference obtained between the original frame set $\mathcal{F}$ and the selected key frame set $\tilde{\mathcal{F}}$, and (b) the processing time reduction achieved for the following 3D reconstruction phase based on these two frame sets with different cardinalities.

In our experiments we utilize the public geo-crowdsourced UGV from *GeoVid* and MediaQ [23]. We retrieved a video dataset as well as its corresponding geo-sensor dataset recorded in two cities, Los Angeles and Singapore. Our experimental dataset contains 345 videos and 77,642 frames. The average length of each video is 55 seconds. Various mobile devices were used for video recording, including the Motorola Milestone, HTC EVO 3D, Samsung Galaxy S4, Asus Transformer and Google Nexus. The video resolution is set to 720 $\times$ 480 by the app. We randomly selected 10 target objects (two in Singapore and eight in Los Angeles) to which we applied our active key frame selection method before the 3D reconstruction.

4.1. Geographic Coverage Gain

In order to evaluate whether our proposed algorithm is able to obtain a minimal coverage gain difference, namely a coverage gain close to the one achieved with the whole frame set, we implemented another heuristic key frame selection strategy based on geographic analysis for comparison.

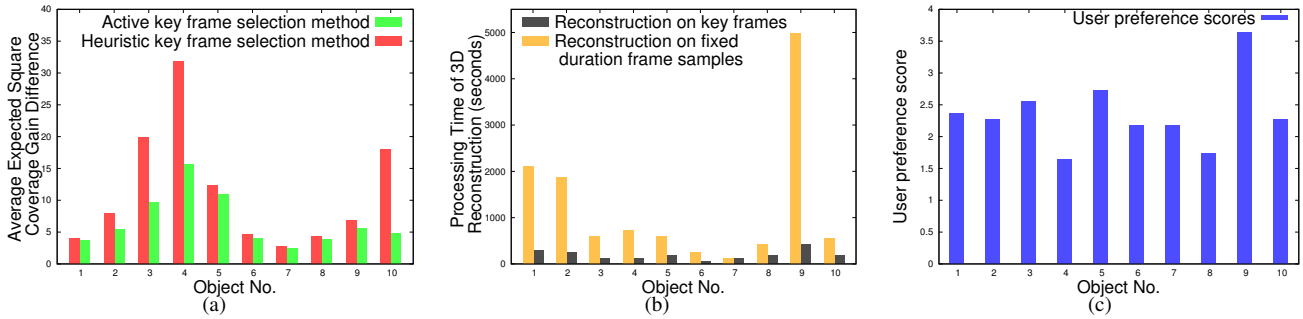
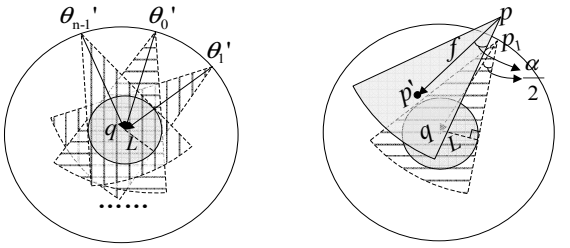


Fig. 2. (a) Average expected square coverage gain difference of 10 target objects. (b) Execution time of target object's 3D reconstruction process. (c) Quality comparison between two 3D reconstruction results on two frame sets for 10 target objects.

Intuitively, in the 2D space of the World Geodetic System (usually an aerial view), for the frames with the same viewing direction towards the target object, we only select the ones in which the target object occupies the largest part of the field-of-view. In other words, from a pixel domain perspective, from the same viewing direction we choose the frame in which the target appears dominantly in the image. In this way we can theoretically extract the most diverse viewing directions with a fixed number of frames. However, in a practical implementation, the "*same viewing direction*" needs to be quantified, which might be all the frames within a certain degree range. Thus, this method has difficulty to achieve a globally optimal solution.



(a) Divided direction groups.

(b) Computation for the direction-location combined score.

Fig. 3. Illustration of the heuristic method.

The heuristic method is designed for comparison and it uses a filter-refine paradigm. It first filters out all video frames that do not capture the target based on their θ value in the accompanying sensor data, which is identical to the step we performed in Equation 1. In the refinement step, as illustrated in Figure 3(a), we equally divide 360 degrees into n directions around the given object q and partition the frame set into n groups based on the camera viewing directions θ. For each group, we select the most geographically covered frame $f\langle \boldsymbol{p}, \theta, \alpha \rangle$ determined by a linear combination score of the distance and the direction difference:

$$I_{score}(f, q) = \beta \times \frac{Dist(q, p')}{MaxDist} + (1 - \beta) \times (1 - \cos(\theta'_j, \theta))$$

As shown in Figure 3(b), the point p' is obtained with a translation by the euclidian distance $Dist(p, p') = Dist(p_1, q) = \frac{L}{\sin\frac{\alpha}{2}}$ along the viewing direction θ of frame f. $MaxDist$ represents the maximal euclidian distance of pairs of distinct objects in $\mathcal{F}$ for normalization. The cosine $\cos(\theta'_j, \theta)$ is the direction difference between the group direction θ'_j and the viewing direction θ of f. The tuning parameter β adjusts the balance between the camera location distance and the direction difference. Finally, in each group, the highest scored frame is extracted into the key frame set. We consider frames within a 10-degree range as belonging to the same viewing angle bin; therefore n is set to 36 and β to 0.2 in our experiments.

In each experiment, we selected at most n frames as a subset (the number of frames may be less due to the absence of coverage from a certain direction in the crowdsrouced UGVs). We set t constant to ensure that the key frame sets have the same cardinality for two methods. We set the remaining parameters at the same values as tested in MAED [22]. Figure 2(a) illustrates the average expected square coverage gain difference, *i.e.*, G_{diff}, calculated between the key frame subset and the whole frame set. The red bar indicates the difference obtained by the key frames selected with the heuristic method. The green bar is the result of $\tilde{\mathcal{F}}$ extracted by our proposed algorithm. Considering all the ten objects of the experiments, our active selection method consistently achieves less difference, in other words, it is successful in finding a subset that achieves a very close coverage of the target object in geographic space compared to the whole frame set. This result is not surprising since our approach guarantees a mathematically minimal solution. Moreover, for some objects such as No. 3 and No. 4, the gain difference is notably decreased. Figure 4 shows two comparison results in detail. We plot the camera locations of the selected key frames in an aerial view. As illustrated, the key frame set selected by our method includes much more viewing directions towards the first object than the heuristic method (Figure 4(a) and Figure 4(b)). For the second object, by contrast, the subsets of the two approaches overlap to a large degree (Figure 4(c) and Figure 4(d)). On average, compared with the heuristic ap-

proach, our key frame selection algorithm decreases the expected square coverage gain difference by 41.93%.

4.2. 3D Reconstruction Performance

We conduct the 3D reconstruction as follows. First, from the frame dataset we extract features with the SIFT method from the VLFeat library. Afterwards, feature matching and bundle adjustment are performed with the SfM bundler library [24]. Next the output of the SfM step is feed into CMVS to divide the image set into clusters of manageable size and allow them to be processed independently and in parallel [3]. Eventually the PMVS2 software is executed to produce a set of oriented points instead of a polygonal (or a mesh) model, where both the 3D coordinates and the surface normals are estimated at each oriented point [25].

We performed all measurements on a 3.4 GHz Intel Core i7-2600 CPU with 4 cores and 8 GB memory. Figre 2(b) illustrates the execution time of the whole 3D reconstruction process for each object. In order to show the efficiency and effectiveness of our active key frame selection method, we sample the collected UGV frames at a fixed duration (1 second in this experiment) as a comparison. The black bar indicates the processing time based on our extracted key frames $\tilde{\mathcal{F}}$, which is significantly less than the processing time based on frame samples described in orange bars. On average, the reconstruction time is shortened by around 17.5 minutes and the maximal number of frames in $\tilde{\mathcal{F}}$ is only 30.

Since we currently do not have a groundtruth 3D model for the reconstruction evaluation, we conducted a user study to compare the quality of the dense point cloud results based on two frame sets. The participants were requested to carefully examine the 3D scene results visualized via MeshLab[1]. For each target object, two results built from a fixed duration sampled frame set and $\tilde{\mathcal{F}}$ were presented, respectively. After judiciously comparing two 3D scene results, participants were asked to provide marks on the quality for both of them (4 – the quality of the reconstruction result based on $\tilde{\mathcal{F}}$ is much better, 0 – the quality of the reconstruction result based on fixed duration frame samples is much better). Twenty-two people participated in this study: 10 males and 12 females, including students, professionals, research staff and faculty. They were asked to consider the completeness of the target object and whether artifacts were visible. We hid the frame set sources of the two results to ensure an unbiased comparison. Figure 2(c) summarizes the results of the user study. As shown, most reconstruction results based on two frame sets were almost the same quality (scores near 2). The key frame set using the active selection method performed slightly worse on objects No. 4 and No. 8 (scores below 2). On the other hand, for some objects such as No. 9, the reconstruction results based on $\tilde{\mathcal{F}}$ were much better than the other frame set. Some 3D reconstruction results from the two frame sets are shown in Figure 5.

[1] meshlab.sourceforge.net/

5. CONCLUSIONS AND FUTURE WORK

We presented a key frame selection method for 3D model reconstruction based on geospatial properties of crowdsourced UGVs. The concept of geographic coverage gain was introduced and the gain difference between the original frames and the key frames was minimized. Our method also incorporates a Manifold Adaptive Kernel to reflect the underlying geometry. Therefore, the key frame set extracted by our algorithm maintains the best coverage of the target object in geographic space. The experimental results demonstrate both the effectiveness and efficiency of our approach. Due to pervasive trends and scalability advances in processing contextual data, key frame selection based on geo-sensor data analysis is practical and can complement a content-based approach. In our future work we plan to combine visual features and more sensors to help with frame extraction.

Acknowledgment

This research has been supported in part 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). This research has also been funded in part by Award No. 2011-IJCX-K054 from the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice, as well as NSF grant IIS-1320149, the USC Integrated Media Systems Center (IMSC) and unrestricted cash gifts from Google and Northrop Grumman. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the sponsors such as the National Science Foundation or the Department of Justice.

6. REFERENCES

[1] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M Seitz, "Multi-view Stereo for Community Photo Collections," in *International Conference on Computer Vision (ICCV)*, 2007.

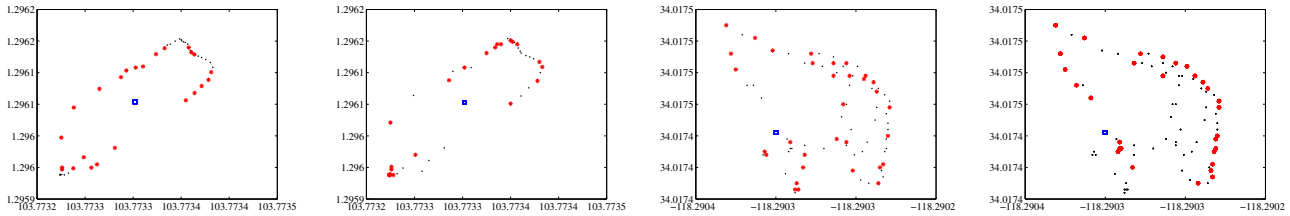
[2] Maxime Lhuillier and Long Quan, "A Quasi-dense Approach to Surface Reconstruction from Uncalibrated Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[3] Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski, "Towards Internet-scale Multi-view Stereo," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[4] Li Ling, Ian S Burrent, and Eva Cheng, "A Dense 3D Reconstruction Approach from Uncalibrated Video Sequences," in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2012.

[5] A Roy Chowdhury, Rama Chellappa, S Krishnamurthy, and T Vo, "3D Face Reconstruction from Video Using a Generic Model," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2002.

[6] Mirza Tahir Ahmed, Matthew N Dailey, Jose Luis Landabaso, and Nicolas Herrero, "Robust Key Frame Extraction for 3D Reconstruction from Video Streams," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2010.



(a) (b) (c) (d)

Fig. 4. Illustration of key frame selection results with two target objects in aerial view. X- and Y-axis denote latitude and longitude. (a) and (b) are the selection results of object No. 1 with our proposed algorithm and the heuristic method, respectively. (c) and (d) are the selection results of object No. 2 with the two approaches. Black and red points indicate the geo-locations of $\mathcal{F}$ and $\tilde{\mathcal{F}}$, respectively. The blue square indicates the geo-location of the target object.

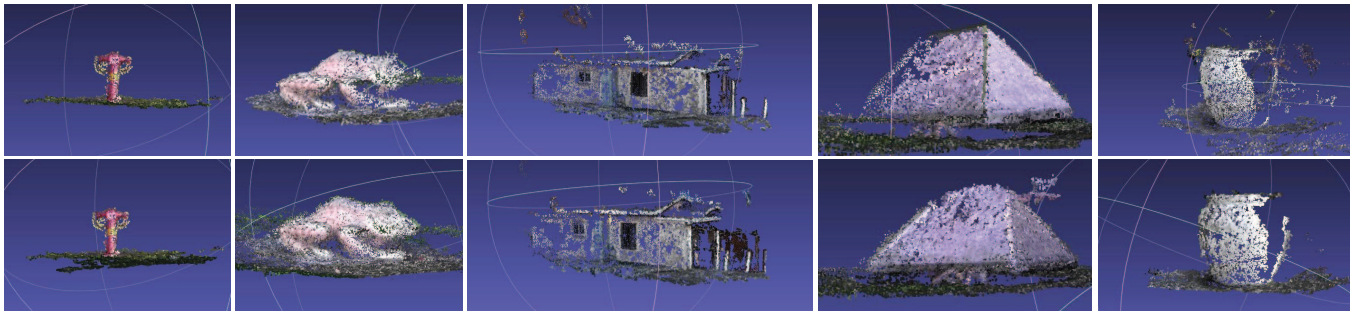


Fig. 5. Illustration of 3D reconstruction results of five target objects. The top row shows the reconstructed 3D models based on the fixed duration sampled frame set. The bottom row shows the results based on our active key frame selection.

[7] Yung-Ho Seo, Sang-Hoon Kim, Kyoung-Soo Doo, and Jong-Soo Choi, "Optimal Keyframe Selection Algorithm for Three-dimensional Reconstruction in Uncalibrated Multiple Images," *Optical Engineering*, 2008.

[8] Jung Kak Seo, Sang Hoon Kim, Cheung Woon Jho, and Hyun Ki Hong, "3D Estimation and Key-Frame Selection for Match Move," in *International Technical Conference on Circuits Systems, Computers and Communications (ITC-CSCC)*, 2003.

[9] Philip HS Torr, "Geometric Motion Segmentation and Model Selection," *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 1998.

[10] Philippos Mordohai, Jan-Michael Frahm, A Akbarzadeh, Brian Clipp, C Engels, David Gallup, Paul Merrell, C Salmi, Sudipta Sinha, B Talton, et al., "Real-time Video-based Reconstruction of Urban Environments," *ISPRS Working Group*, 2007.

[11] Sakire Arslan Ay, Roger Zimmermann, and Seon Ho Kim, "Viewable Scene Modeling for Geospatial Video Search," in *ACM International Conference on Multimedia*, 2008.

[12] Seon Ho Kim, Sakire Arslan Ay, and Roger Zimmermann, "Design and Implementation of Geo-tagged Video Search Framework," *Journal of Visual Communication and Image Representation*, pp. 773–786, 2010.

[13] Guanfeng Wang, Beomjoo Seo, and Roger Zimmermann, "Motch: An Automatic Motion Type Characterization System for Sensor-rich Videos," in *ACM International Conference on Multimedia*, 2012.

[14] Richard K Burkhard, *Geodesy for the Layman*, US Department of Commerce, National Oceanic and Atmospheric Administration, 1985.

[15] Kai Yu, Jinbo Bi, and Volker Tresp, "Transductive Experiment Design," 2005.

[16] Luming Zhang, Yue Gao, Ke Lu, and Jialie Shen, "Representative Discovery of Structure Cues for Weakly-Supervised Image Segmentation," *IEEE Transcations on Multimedia*, 2014.

[17] Luming Zhang, Yue Gao, Rongrong Ji, Qionghai Dai, and Xuelong Li, "Discovering Active Viewing Paths for Semantics-Aware Photo Cropping," *IEEE Transcations on Image Processing*, 2014.

[18] Stephen Poythress Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[19] Kai Yu, Jinbo Bi, and Volker Tresp, "Active Learning via Transductive Experimental Design," in *International Conference on Machine Learning*, 2006.

[20] Kai Yu, Shenghuo Zhu, Wei Xu, and Yihong Gong, "Non-greedy Active Learning for Text Categorization Using Convex Ansductive Experimental Design," in *International ACM SIGIR Conference*, 2008.

[21] Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin, "Beyond the Point Cloud: from Transductive to Semi-supervised Learning," in *International Conference on Machine Learning*, 2005.

[22] Deng Cai and Xiaofei He, "Manifold Adaptive Experimental Design for Text Categorization," *IEEE Transactions on Knowledge and Data Engineering*, 2012.

[23] Seon Ho Kim, Ying Lu, Giorgos Constantinou, Cyrus Shahabi, Guanfeng Wang, and Roger Zimmermann, "MediaQ: Mobile Multimedia Management System," in *5th ACM Multimedia Systems Conference*, 2014, pp. 224–235.

[24] Noah Snavely, Steven M Seitz, and Richard Szeliski, "Photo Tourism: Exploring Photo Collections in 3D," in *ACM Transactions on Graphics*, 2006.

[25] Yasutaka Furukawa and Jean Ponce, "Accurate, Dense, and Robust Multiview Stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.