Vol.2015-CVIM-197 No.35
2015/5/19

Four-dimensional City Modeling using Vehicular Imagery

Ken Sakurada[1,a)] Takayuki Okatani[1,b)]

Abstract: We propose a novel method for four-dimensional city modeling using vehicular imagery. Motivation behind this study is to estimate and visualize damage and recovery process of tsunami-damaged area. To conduct the research, we have been recording the images in tsunami-damaged area periodically driving a car on which omnidirectional camera is mounted. To estimate temporal changes of a wide area using vehicular imagery, there are many challenges to overcome, for example, limitation of camera viewpoint, coverage of vehicular imagery and illumination condition. Furthermore, a 3D model is not always available for every city, and the scene images do often not have sufficient visual features to perform precise registration. In the case of a wide-area disaster, it is computationally prohibitive to reconstruct the three-dimensional structure of entire areas. To overcome these difficulties, we propose 2D, 3D and object-based change detection methods. The 2D method detects scene change from an image pair using visual features of convolutional neural network. The 3D method estimates structural changes of scene from images taken at multiple viewpoints even if there is depth ambiguity of a scene. The object-based method estimates land surface condition of an entire city integrating aerial and street-view imagery, which are taken at vastly different viewpoints. The experimental results show that our methods can accurately and effectively estimate temporal changes of a city.

1. Introduction

On March 11th, 2011, Great East Japan Earthquake brought catastrophic damage to the north east of Japan. The earthquake centered at 70 km offshore of Miyagi prefecture and recorded the magnitude of 9.0. The Tsunami caused by the earthquake reached 40.1 meters height at maximum. The earthquake caused giant Tsunami whose maximum height was 40.1 meter and the Tsunami gave serious damages to the Pacific coast area of the Tohoku. The great earthquake, the giant Tsunami and the aftershocks caused landslide disaster, fire, land subsidence and ground liquefaction. The secondary disasters spread to a very wide area including Fukushima prefecture where first nuclear power plant resulted in the release of radioactive substances after the power loss caused by the Tsunami. The earthquake triggered an all-time wide-area complex disaster.

Accurate understanding of damage and temporal scene change is important to reduce secondary damage and quick recovery and restoration. Aerial image is one of the most frequently used sensory information to investigate a wide area damaged by a disaster. For example, it is possible to observe the land-surface condition using a satellite image of visible light, and estimate an area condition such as inundation using aerial image of infrared light and microwave both during day and night. However, aerial image has some drawbacks, such as low resolution and large variation of illumination condition due to weather change, since aerial image observes the ground from high-altitude in the sky. Furthermore, there are many areas invisible from the sky due to coverage by a roof, an elevated road and a pedestrian bridge (Fig.1). Street-view image is essential to supplement such missing observation from aerial image and understand the detail of city condition.

Fig. 1 Areas unseen from a satellite

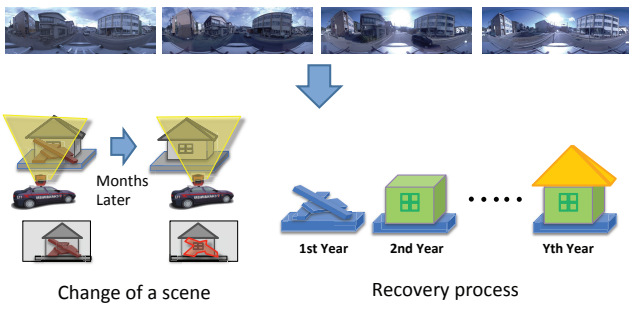


Fig. 2 Overview of 4-dimensional city modeling using vehicular imagery

The objective of this study is to visualize the damage and recovery/restoration process of tsunami-damaged area. About one month after the earthquake, we started recording the damages and recoveries of tsunami-damaged areas driving a car on which an omni-directional camera and a GPS receiver are mounted. The time interval of the recording is from 2 to 6 months depending on the recovery progress of the areas. The target area is the costal area of almost 500 kilometer which observed serious Tsunami-damages caused by Great East Japan Earthquake in 2011 (from Aomori to Fukushima prefecture). We have recorded about 40 terabytes of image data so far.

[1] Tohoku University, 6-6-01 Aramaki Aza Aoba, Aoba-ku, Sendai-shi, Miyagi, Japan
[a)] sakurada@vision.is.tohoku.ac.jp
[b)] okatani@vision.is.tohoku.ac.jp

Figure 2 describes the overview of 4-dimensional city modeling using vehicular imagery. The image archival activity has been periodically recording the scenes of the cities in the tsunami-damaged areas. From the periodical observation, the 4D modeling method detects scene change, and estimates city-scale and regional-scale temporal changes. The simplest method for estimating temporal change is to directly differentiate results of three dimensional reconstruction. However, to estimate temporal changes of regional-scale area using vehicular imagery, there are challenges to overcome. First, vehicle-mounted camera only captures the scene alongside street. The limited view points causes the depth ambiguity and makes it difficult to densely reconstruct the 3D structure of the scene. Second, vehicle image cannot cover occluded areas and unreachable areas. A single vehicle image has limited physical range. Third, regional-scale change detection requires too much computational resources since one city has several thousands to several tens of thousands of image pairs, especially for 3D reconstruction and pixel-level registration.

The strategy for this paper to overcome these challenges is as follows. First, the proposed method roughly but quickly detects 2D scene change of entire areas from an image pair. Next, the method detects accurate structural change where detailed analysis is necessary. Finally, the method estimate city-scale temporal change. This paper proposes a novel method for 2D change detection, structural change detection and city-scale land surface condition analysis.

Section 2 Digital Image Archive of Tsunami-damaged Area This section describes the image dataset built for studying 4D city modeling. The archival process started since about one month after the Japan earthquake of March 11, 2011, and accumulated about 40 TB of data. The proposed methods estimate the recovery process of the tsunami-damaged area using this image dataset.

Section 3 Three-dimensional Reconstruction As preliminary study, this section shows the results of sparse and dense 3D reconstructions. For sparse reconstruction, a standard Structure from Motion (SfM) is performed which is extended to omnidirectional image. Using the camera poses of the SfM, Patch-based Multi-view Stereo (PMVS2) [15] generates dense city-models. Furthermore, this section shows the result of temporal change detection that naively compares the reconstructed structure over time. The results show that the naive method is not enough to understand the detail of city condition. To overcome this challenge, this paper proposes the following three methods for 4D city modeling.

Section 4 2D Change Detection This section describes the method to detect 2D scene changes from an image pair using grid feature. Several previous approaches of change detection require 3D model of a scene and pixel-level registration between different time images. In the case that 3D model is not available, it is difficult to directly apply the previous methods to the change detection problem. Furthermore, it is computationally prohibitive to estimate scene change of wide area using 3D model and pixel-level registration. The proposed method can detect scene change without pixel-level registration integrating convolutional neural network (CNN) feature with superpixel segmentation. The method can reduce the computational time and detect change of entire tsunami-damaged areas. The experimental results show that the proposed method effectively integrates high discrimination of CNN feature and accurate segmentation of superpixel.

Section 5 3D Change Detection This section describes a method for detecting temporal changes of the three-dimensional structure of an outdoor scene from its multi-view images captured at two separate times. The method estimates scene structures probabilistically, not deterministically, and based on their estimates, the method evaluates the probability of structural changes in the scene, where the inputs are the similarity of the local image patches among the multi-view images. The proposed method is compared to the approach that use multi-view stereo (MVS) to reconstruct the scene structures of the two time points and then differentiate them to detect changes. The experimental results show that the proposed method outperforms MVS-based methods.

Section 6 Land Surface Condition Analysis This section presents a unified framework for robustly integrating image data taken at vastly different viewpoints to generate large-scale estimates of land surface conditions. To validate the proposed approach, this study attempts to estimate the amount of post-tsunami damage over the entire city of Kamaishi, Iwate Prefecture (over 4 million square-meters). The results show that the proposed approach can effectively integrate both micro and macro-level images, along with other forms of meta-data, to effectively estimate city-scale phenomena. Experiments evaluate the proposed approach on two modes of land condition analysis, namely, city-scale debris and greenery estimation, to show the ability of the proposed method to generalize to a diverse set of estimation tasks.

Section 7 Conclusion The paper concludes with a summary, and discusses a consideration of future extensions of this work, including open and remaining questions.

1.1 Related Work

This section will review the previous work relevant to understanding 4-dimensional city modeling in terms of temporal change detection and city-scale analysis.

1.1.1 City Modeling

The problem of measuring and documenting a city is the objective of photogrammetry, remote sensing and computer vision community [6], [17], [23], [47], [48]. City modeling is, for example, 3D reconstruction, land-use mapping and scene change estimation. There are many input data types to reconstruct a city other than image, for example, light detection and ranging (LiDAR), digital elevation map (DEM), digital terrain model (DTM) and digital surface model (DSM). The followings focus on automatic methods using image and LiDAR. For the method using other data sources and interactive methods, please refer to the paper [37].

There are multiple types of devices to measure a city, for example, digital camera and LiDAR mounted on mobile devices or systems such as smartphone, vehicle, UAV, airplane and satellite. Snavely et al. propose a method to reconstruct an entire city using unstructured images which were captured from a variety of view points using mobile devices and uploaded on the Internet [1]. Pollefeys et al. proposed an approach for dense 3D recon-

struction from unregistered Internet-scale photo collections with about 3 million of images within the span of a day on a single PC [14]. Furthermore, Pollefeys developed a system for automatic, georegistered, real-time 3D reconstruction from video of urban scenes [40].

Poullis and You proposed a method for massive city-scale reconstruction using imagery and LiDAR [44]. This system automatically creates lightweight, watertight polygonal 3D models from LiDAR data captured by an airborne scanner[41], [42], [43], [44]. The technique is based on the statistical analysis of the geometric properties of the data, which makes no particular assumptions about the input data. Zhou and Neumann proposed a similar approach [69], [70]. Lafarge and Mallet developed a method for modeling cities from 3D-point data providing a more complete description than existing approaches by reconstructing simultaneously buildings, trees and topologically complex grounds [28], [29]. Cabezas et al. proposed an integrated probabilistic model for multimodal fusion of aerial imagery [4], LiDAR data and GPS measurements. The model of their method allows for analysis and dense reconstruction (in terms of both geometry and appearance) of large 3D scenes. One of its advantages is that it explicitly models uncertainty and allows for missing data. This work takes the advantages of the city modeling methods.

1.1.2 Temporal Change Detection

Many researchers have worked on temporal change detection of a scene. However, most of them consider the detection of 2D changes (i.e., those only in image appearance), whereas the objective of this study is to detect changes in 3D structure of scenes.

The standard problem formulation of 2D change detection [39], [45] is an appearance model of a scene is learned using its n images and then based on $n + 1^{st}$ image, it is determined whether a significant change has occurred. Most of the studies of 3D change detection [8], [24], [25], [39], [58] follow a similar formulation; namely, a model of the scene in a "steady state" is built and a newly-captured image(s) is compared against it to detect changes.

In [39], targeting at aerial images capturing a ground scene, Pollard and Mundy proposed a method that learns a voxel-based appearance model of a 3D scene from its 20–40 images. Crispell et al. later improved method to minimize storage space is presented in [8]. In [25], Ibrahim and David proposed a method that detects scene changes by estimating the appearance or disappearance of line segments in space. All of these studies create an appearance model of the target scene from a sufficiently large number of images, unfortunately, this approach does not work due to lack of images. Such an approach is appropriate for aerial or satellite imagery or the case of stationary cameras, but is not appropriate for the images taken in our setting.

The alternative approach is to obtain a 3D model of the scene from other sensors or methods than the images used for the change detection. In [24], assuming that the 3D model of a building is given, the edges extracted in its aerial images are matched with the projection of the 3D model to detect changes. The recent study of Taneja et al. [58] is of the same type. Their method detects temporal changes of a scene from its multi-view images, and thus it is close to ours from an application point of view. However, their motivation is to minimize the cost needed for updating the 3D model of a large urban area, and thus, they assume that a dense 3D model of the target scene is given.

The proposed method in this paper differs from all of the above in the formulation of the problem. In the proposed formulation, the changes of a scene are detected from two sets of images taken at two different time points. The two image sets are "symmetric" in a sense that they have similar sizes and are of the same nature. The proposed method does not assume that a dense 3D model of the scene is given, or created from the input images themselves, as it is difficult for the images captured from a ground vehicle-mounted camera. If the dense model is required, it is necessary to have a large number of multi-view images captured from a variety of viewpoints [1], [7], [40], [54], [67], [68], or to use a range sensor.

In the sense that the input data are symmetric, the proposed method might be close to the study of Schindler and Dellaert [52]. They propose a method that uses a large number of images of a city that are taken over several decades to perform several types of temporal inferences, such as estimating when each building in the city was constructed. However, besides the necessity for a large number of images, their method represents scene changes only in the form of point clouds associated with image features.

1.1.3 City-scale Surface Condition Analysis

There has been significant advances in the state-of-the-art techniques for quantitative geometric interpretations of large-scale city scenes. Methods for city-scale 3D reconstruction have been proposed using thousands of images gathered from Internet images [1], [54]. Similar techniques have been proposed for images captured by a vehicle-mounted camera [40], [61] or aerial images [33], [57], [66]. Street-view images have also been combined with aerial images for the purpose of improving 3D reconstruction, where 3D point clouds have been projected to the ground plane and aligned with edges of buildings detected from aerial images [26] or building maps [56]. There has also been work using aerial and street view images taken several months or decades apart [24], [39], [45], [52], [59] to understand temporal changes of a scene. The focus of these previous approaches are on a quantitative geometric interpretation of the scene where local visual features are matched directly to estimate camera pose using epipolar geometry [20]. This work aims to push beyond a purely geometric understanding of the scene towards a more qualitative understanding of city conditions. For instance, the aim is not only to estimate the 3D geometry of a building but also the condition of the building or the condition of the ground surrounding a building.

There also has been work focused on the qualitative estimation of land condition over large-scale environments. In the field of remote sensing, coarse land surface conditions have been estimated using aerial color images, aerial infrared light and aerial microwave sensing [31], [35], [36], [53], [64], [65]. Color aerial images have been applied to land condition estimation for vegetation monitoring [3], [10], [18], land cover mapping, and flood risk and damage assessment [22], [63]. For example, forest maps [16], [19], [55] are an important source of information for monitoring and reducing deforestation, allowing environmental scien-

tists to know how forested areas increase or decrease in over the entire earth.

Apart from aerial imaging using color cameras, many other modes of sensing have been proposed for estimating coarse large-scale land surface conditions. Digital elevation map (DEM) [16], Spectroradiometer (MODIS), high resolution radiometer (AVHRR) and Synthetic Aperture Radar (SAR) have been proposed to improve accuracy of estimating large-scale land surface condition. However the resolution of satellite-mounted MODIS and AVHRR only measure surface conditions over a very rough resolution – typically over a cell size of a several hundred meters. As such, these works do not utilize street-level sensing which are too detailed for their estimation task. However, this work aims at estimating land conditions on a cell size closer to 20 meters wide.

The proposed work fills a void between detailed geometric reconstructions of city-scale structures and coarse qualitative estimation of land conditions. The proposed method uses known techniques to provide an accurate geometric model of the city and use state-of-the-art object recognition results carefully registered to the scene geometry to understand the qualitative conditions of the entire city.

2. Tsunami Damage Archive

This chapter discusses about the detail of the image dataset used in this research. We have been recording images in tsunami-devastated areas using a vehicle mounted camera since about one month after the earthquake. This image dataset consists of city-scale street-view images of different times. This research proposed some methods estimating city-condition and temporal change from the dataset.

2.1 Image Acquisition

Since about one month after the earthquake, we started recording the damages and recoveries of these areas mainly using a vehicle-mounted omni-directional camera (Fig. 3).

The image archive activity is periodically acquiring the images of the tsunami-devastated areas in the northern-east coast of Japan. The images are captured by a vehicle having an omni-directional camera (Ladybug3 of Point Grey Research Inc.) on its roof. An image is captured at about every 2m on each city street to maintain the running speed of the vehicle under the constraint of the frame rate of the camera.

2.1.1 Measurement Vehicle

Figure 3 shows our measurement vehicle which mounts an omni-directional camera (Ladybug3 or Ladybug5 of Point Grey Research Inc.) and a receiver of Differential Global Positioning System (DGPS) (R100 of Hemisphere Inc.). A Ladybug camera has six CCD image sensors. Figure 4 shows image of each camera of Ladybug. Using these raw images, computational photography method can generate omnidirectional panoramic image, perspective image of arbitrary view-direction and image of dome projection. In this research, Structure from Motion (SfM) uses the panoramic image and recognition methods use perspective image. Our approach uses perspective image cropped in the left or right direction since images in the left and right direction have rich information of city scene.

Fig. 3 Measurement vehicle equipping an omnidirectional camera (Ladybug 3) and GPS.

Fig. 4 Images of each camera of a Ladybug camera.

2.1.2 Measured area

Figure 5 shows the measured area period of this image archive activity. In the first one month, this activity mostly covered the entire devastated areas across the three prefectures whose total length is almost 400 kilometers. Figure 6 shows periodically measured area in Kamaishi. The color line shows a trajectory of our measurement vehicle. Different color shows different time data. The blue circles show the area where ordinary people could enter because of recovery operations one year after the tsunami. It takes about two weeks to measure the entire devastated areas across the three prefectures. We have gotten about 40 terabytes of image data until December, 2014.

This image archive activity is different from similar activities conducted by other parties such as Google Inc. in that the goal of this activity is to record the temporal changes of these areas and thus we have been periodically recorded these areas.

2.2 Temporal changes

Figure 7 shows the examples of panoramic images which we periodically captured in the tsunami-devastated areas. It is possible to understand from these images that there are temporal changes. For example, big damages due to tsunami and recovery operations. However, it is not easy to understand damage and recovery process of an entire city only by looking at these images. Furthermore, these images have differences of viewpoint and illumination condition between different time data. The proposed method of this paper enables it to automatically estimate and visualize temporal change of an entire city using street-view images and other metadata.

3. 3D Reconstruction

This section explains methods to reconstruct three-dimensional structure of a scene using a sequence of omnidirectional panoramic images and to estimate temporal changes using the reconstruction results.

The simplest baseline for estimating temporal change is to directly differentiate results of three dimensional reconstruction. To

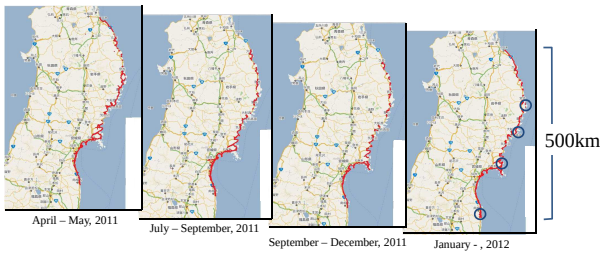


Fig. 5 Area and period of our image archive activity.

Fig. 6 Measured area in Kamaishi. The lines show the trajectories of the measurement vehicle. Different color shows different time data.

Fig. 7 Example of temporal changes in tsunami-devastated areas.

differentiate different time data, it is necessary to align the data in a common coordinate. Later, section 5 compares this baseline against the proposed approach.

3.1 Structure from Motion

Structure from Motion (SfM) is a general method to estimate camera pose using images [1], [40]. As mentioned in section 2, the image archive activity is capturing sequential omnidirectional images in tsunami-devastated areas. Hence, in this paper, SfM estimates camera pose using 360 degrees field of view panoramic images [61].

The method is summarized as follows:

(1) Feature points are extracted with the Speed Up Robust Feature (SURF) [2] and tentative matching is obtained for two consecutive images using the descriptors of these feature points.
(2) Essential matrices are calculated with the five point algorithm [38]. At that time, mismatches of feature points are rejected using Random Sample Consensus (RANSAC) [12].
(3) Camera poses and positions of feature points are calculated using those essential matrices.
(4) Camera poses and position of 3D points are optimized to minimize reprojection errors of feature points.

3D point clouds of which each point has an image descriptors is generated through (1)-(4) processes.

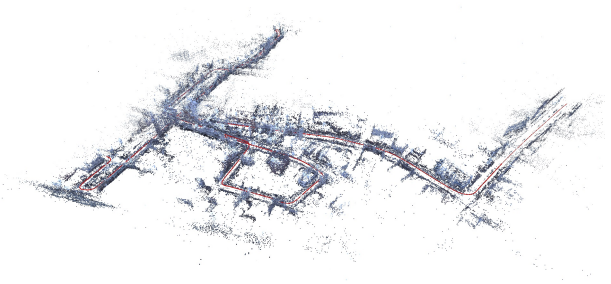


Fig. 8 Example of city-scale 3D reconstruction (April, 2011, Kamaishi, Iwate). This result consists of one thousand omnidirectional images.

Figure 8 shows a result of city-scale three-dimensional reconstruction using one thousand omnidirectional images (April, 2011, Kamaishi, Iwate). Red line shows trajectories of a camera (i.e. our measurement vehicle). The point clouds show structural objects, such as building, telegraph pole, and tree. This reconstruction result shows the structure of the entire city. However, it is not easy to understand the detail of the structure due to the sparseness of feature points.

The left of figure 9 shows reconstruction results consisting of sparse feature points using images captured at a same location in July, 2011. The right of figure 9 show dense reconstruction results using Patch-based Multi-view Stereo (PMVS2) [15] corresponding to the left of Fig. 9. The sparse reconstruction results represent the entire shape of the street well. And the dense reconstruction results using PMVS2 represents the detail of the scenes well although they have some lacks of the structures, especially for texture-less area. However, it is difficult to understand the details of the scenes using the sparse results, and, regional-scale 3D reconstruction requires too much computational resources since one city has several thousands to several tens of thousands of image. basically, dense reconstruction methods consume much computational resources.

3.2 Baseline Method for Temporal Change Detection

Baseline method for temporal change detection is to directly compare three dimensional structures of different times. In this

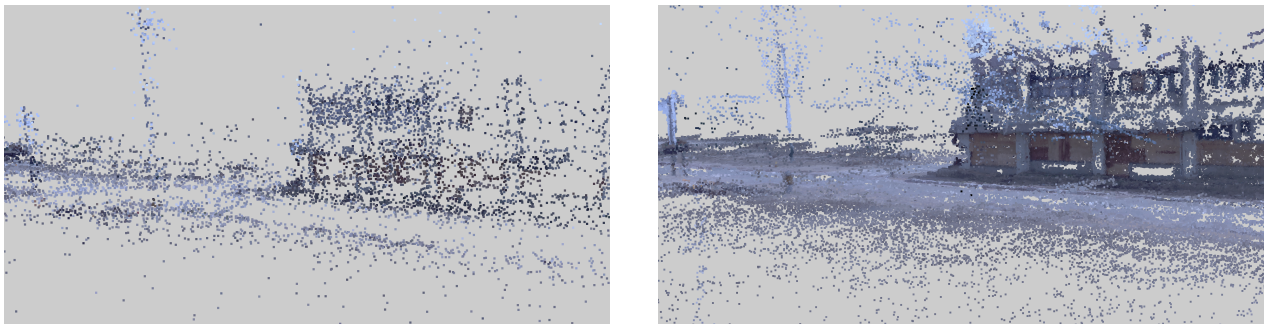


Fig. 9 Example of dense 3D reconstruction scene using sparse feature points (left) and PMVS2 (right) (July, 2011, Rikuzentakata, Iwate).

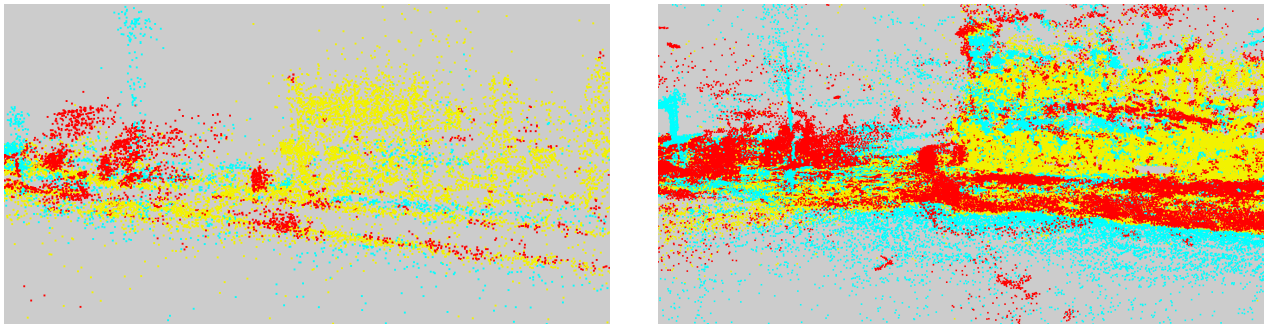


Fig. 10 A scene of change detection using sparse feature points (left) and dense 3D points of PMVS2 (right) comparing April, 2011 and July, 2011, Rikuzentakata, Iwate. Red, blue and yellow show disappearance, appearance and no-change, respectively.

section, as a baseline of temporal change detection, results of temporal change detection based on point clouds are shown.

To differentiate different time structure, it is necessary to register different time data in the common coordinate. The detail of our registration method is described in sec. 5.1. The summary of the method is as follows. First, SfM is performed independently for each sequence. Next, the two reconstructions are roughly aligned with a similarity transform using RANSAC [13]. Finally, bundle adjustment is performed for the extended SfM problem, in which the sum of the reprojection errors for all the correspondences is minimized.

After the alignment, temporal changes is detected by differentiating the two reconstruction. For the change detection based on point clouds, it is necessary to consider difference of point densities because point density reconstructed using SfM is basically in inverse proportional to distances from cameras. Hence, first, the method of this paper calculates the average distance d_{same} between the point and the nearest N points of the same time data, and the average distance d_{diff} between the point and the nearest N points of the other time data. The point is labeled as "Change" if $d_{\text{diff}} > 2d_{\text{same}}$, "Not Change" otherwise. If the point is observed in only old or new data, the point is labeled as "Dissapeared" and "Appeared", respectively.

The left of figure 10 shows the results of change detection comparing sparse reconstruction results of April, 2011 and July, 2011, Rikuzentakata, Iwate. Red, blue and yellow show disappearance, appearance and no-change, respectively. These figures show important changes of the scene, for example, debris along the street were removed (red), and telegraph poles were built (blue) in an early stage of the recovery operation. Some ground areas are labeled as "Appeared" because those areas are occluded by debris in the images of April.

The right of figure 10 show the results of change detection comparing dense reconstruction results using PMVS2. The results of change detection using dense reconstructions show the detail of the changes well, especially for texture-less areas (e.g. building wall, the ground).

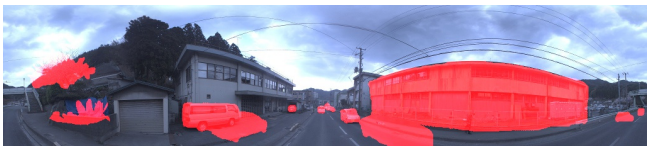
However, even dense change detection results have some missing parts due to the ambiguity of the estimated scene depth. For getting accurate shape of a scene change, it is necessary to maximize the usage of image information. Our probabilistic method of change detection is explained in Sec. 5.

4. 2D Change Detection

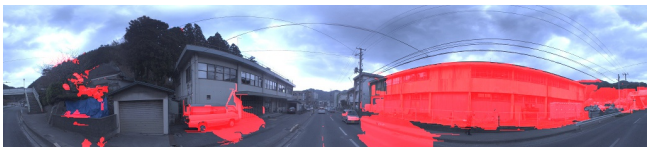
This section considers a problem of detecting scene change from a pair of images taken at different time. The goal behind this study is to estimate city-scale scene change of relatively short term due to disaster, for example, earthquake and tsunami. Understanding of scene change only by driving a vehicle is effective for disaster reduction, quick recovery and restoration.

However, there are some challenges for estimating scene change using vehicular imagery due to the differences of camera view points, illumination condition, photographing condition, sky (e.g. cloud) and ground (e.g. dust on the road) between different time images. It is necessary to develop a change detection method robust for these difficulties.

Input image (database)



Ground-truth superimposed on input image (query)



Change estimation (binarized)

Fig. 11 Results of change detection using pool-5 feature of CNN (Frame No. 0 of 20).

Input image (database)

Ground-truth superimposed on input image (query)

Change estimation (binarized)

Fig. 12 Results of change detection using pool-5 feature of CNN (Frame No. 1 of 20).

Some previous approaches of change detection assume either or both a 3D model of a scene and pixel-level registration. In the case of wide-area disaster, it is computationally prohibitive to reconstruct three dimensional structure of entire areas and to estimate accurate camera pose.

We developed a novel method to detect scene change without 3D model and pixel-level registration integrating convolutional neural network (CNN) feature with superpixel segmentation. First, the proposed method roughly but quickly estimates scene change of entire areas from image pairs which are aligned using Global Positioning System (GPS) data. Next, the method described in section 5 estimates structural scene change of the areas where detailed analysis is necessary. These two steps enable us to quickly and accurately estimate scene change of wide-area.

Figures 11 and 12 show examples of the results of the 2D change detection. It is observed from them that the proposed method was able to correctly detect the scene changes, for example, demolished and new buildings, cars and debris.

5. 3D Change Detection

This section describes a method for detecting temporal changes of the three-dimensional structure of an outdoor scene from its multi-view images captured at two separate times [50]. The proposed method detects accurate structural change of the areas where the result of 2D change detection requests detailed analysis. The method estimates scene structures probabilistically, not deterministically, and based on their estimates, it evaluates the probability of structural changes in the scene, where the inputs are the similarity of the local image patches among the multi-view images. The aim of the probabilistic treatment is to maximize the accuracy of change detection, behind which there is our conjecture that although it is difficult to estimate the scene structures deterministically, it should be easier to detect their changes. The proposed method is compared with the methods that use multi-view stereo (MVS) to reconstruct the scene structures of the two time points and then differentiate them to detect changes. The experimental results show that the proposed method outperforms such MVS-based methods.

5.1 From image acquisition to change detection

As mentioned earlier, we have been periodically acquiring the images of the tsunami-devastated areas in the northern-east coast of Japan. The images are captured by a vehicle having an omni-directional camera (Ladybug3 or Ladybug5 of Point Grey Research Inc.) on its roof. An image is captured at about every 2m on each city street to minimize the total size of the data as well as to maintain the running speed of the vehicle under the constraint of the frame rate of the camera.

The goal of the present study is to detect the temporal changes of a scene from its images thus obtained at two separate times. Figure 13 shows how the input images are processed. For computational simplicity, our algorithm for change detection takes as inputs not the omni-directional images but the perspective images cropped from them. The algorithm also needs the relative camera poses of these images. To obtain them, we perform SfM for each sequence followed by registration of the two reconstructions, which are summarized below.

The algorithm shown in the next section uses only several perspective images to detect changes of a scene. For the reason of accuracy, however, to obtain their camera poses, we perform SfM and registration not with these perspective images alone but with a more number (e.g., 100 viewpoints) of omni-directional images that contain these viewpoints. To be specific, we do this in the following two steps. First, we perform SfM independently for each sequence. We employ a standard SfM method [20], [34], [62] with extensions to deal with omni-directional images [61]. Next, we register the two 3D reconstructions thus obtained as follows. We first roughly align the two reconstructions with a similar-

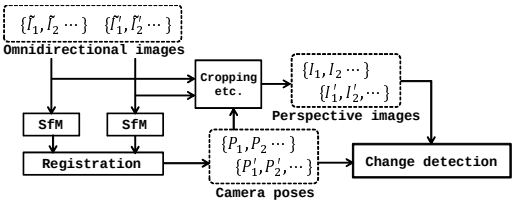
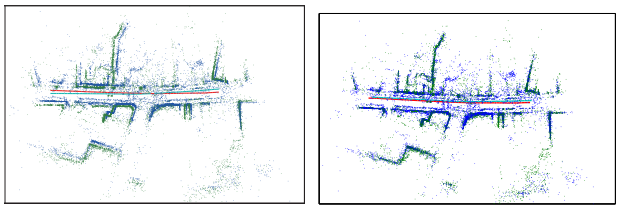


Fig. 13 Data flow diagram.



(a) (b)

Fig. 14 Registration of 3D reconstructions from two image sequences taken at different times. (a) Initial estimate. (b) Final result.

ity transform; putative matches of the feature points are established between the two sequences based on their descriptor similarity, for which RANSAC is performed [13]. For the aligned reconstructions, we reestablish the correspondences of feature points by incorporating a distance constraint. Using the newly established correspondences along with original correspondences within each sequence, we perform bundle adjustment for the extended SfM problem, in which the sum of the reprojection errors for all the correspondences is minimized. Figure 14(a) shows the initial rough alignment of the two reconstructions and (b) shows the final result.

5.2 Detection of temporal changes of a scene

5.2.1 Problem

Applying the above methods to two sequences of omnidirectional images, we have the camera pose of each image represented in the same 3D space. Choosing a portion of the scene for which we want to detect changes, we crop and warp the original images to have two sets of perspective images covering the scene portion just enough, as shown in Fig. 15. In this section, we consider the problem of detecting scene changes from these two sets of multi-view perspective images. For simplicity of explanation, we mainly consider the minimal case where there are two images in each set.

5.2.2 Outline of the proposed method

We denote the first set of images of time t by $\mathcal{I} = \{I_1, I_2\}$ and the second set of time t' by $\mathcal{I}' = \{I'_1, I'_2\}$. As shown in Fig. 16, one of the two image sets, $\mathcal{I}$, is used for estimating the depths of the scene, and the other image set $\mathcal{I}'$ is used for estimating changes of the scene depths. (These may be swapped.) Choosing one image from $\mathcal{I}$, say I_1, which we call a *key frame* here, the proposed method considers the scene depth at each pixel of I_1 and estimates whether or not it changes from t to t'. The output of the method is the probability of a depth change at each pixel of I_1.

For the first image set $\mathcal{I}_1$, its images are used to estimate the depth map of the scene at t. To be specific, not the value of the depth d but its probabilistic density $p(d)$ is estimated. For the

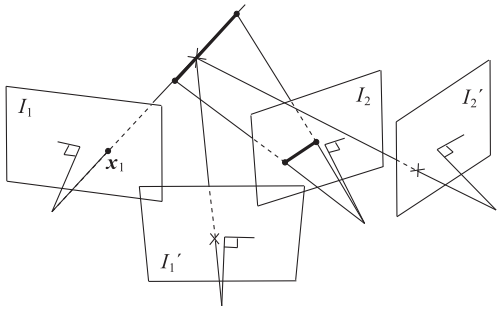


Fig. 15 Geometry of two sets of multi-view perspective images taken at different times. For each pixel $\mathbf{x}_1$ of I_1, the probability that the scene depth has changed is estimated.

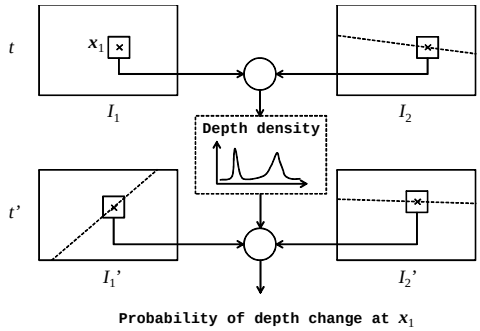


Fig. 16 Outline of the proposed method. The probability density of the scene depth at a point $\mathbf{x}_1$ of I_1 is estimated from I_1 and I_2. This is combined with the comparison of the local patches of I'_1 and I'_2 to estimate the probability that the scene depth changes at $\mathbf{x}_1$ between t and t'. Note that the patches are compared only among the images taken at the same time. The broken lines in the images indicate epipolar lines associated with $\mathbf{x}_1$.

other set $\mathcal{I}'$, a spatial point having depth d at a certain pixel of the key frame I_1 is projected onto I'_1 and I'_2, respectively, as shown in Fig. 15, and then the similarity s'_d of the local patches around these two points is computed. The higher the similarity is, the more the spatial point is likely to belong to the surface of some object in the scene at t', and the inverse is true as well. The similarity s'_d is computed for each depth d, which gives a density function of d that is similar to $p(d)$.

By combining these two estimates, $p(d)$, and s'_d, the proposed method calculates the probability of a depth change. In this process, the change probability evaluated for each depth d is integrated over d to yield the overall probability of a depth change. This makes it unnecessary to explicitly determine the scene depth neither at t nor t'. This is a central idea of the proposed method.

It should also be noted that our method evaluate the patch similarity only within each image set of $\mathcal{I}_1$ and $\mathcal{I}_2$. This makes it free from the illumination changes between the time points of the image capture.

5.3 Experimental results

We conducted several experiments to examine the performance of the proposed method. For the experiments, we chose a few scenes and their images from our archives mentioned in Sec.2. The chosen images are taken at one and four months after the tsunami[*1]. Typically, a lot of tsunami debris appear in the ear-

[*1] The data used in this study (the omni-directional image sequences of the chosen streets and our estimates of their camera poses) are available from

lier images, whereas they disappear in the later ones because of recovery operations. We wish to correctly identify their disappearance in the later images.

The proposed method uses two or more images for each time. In the experiment, we use four images of consecutive viewpoints for each time, i.e., three pairs of images. These are perspective images (cropped from omni-directional images) of 640×480 pixel size. The disparity space is discretized into 128 blocks ($n = 128$). Assuming that there is no prior on the probability of scene changes, we set $p(c = 1) = 0.5$. It is noted, though, that in the experiments, the results are very robust to the choice of this value. These are fixed for all the experiments.

5.3.1 Compared methods

We compared our method with MVS-based ones, which first reconstruct the structures of a scene based on MVS and differentiate them to obtain scene changes. We consider two MVS algorithms for 3D reconstruction, PMVS2 [15] and a standard stereo matching algorithm for it.

In the former case, PMVS2 is applied to a sufficiently long sequence of images (e.g., 100 viewpoints) covering the target scene. Our omni-directional camera consists of six cameras and records six perspective images at each viewpoint. All these six images per viewpoint are inputted to PMVS2 after distortion correction. PMVS2 outputs point clouds, from which we create a depth map viewed from the key frame. This is done by projecting the points onto the image plane in such a way that each point occupies an image area of 7×7 pixels. Two depth maps are created for the two time points and are differentiated to obtain scene changes. We call the overall procedure PMVS2.

In the latter case, a standard stereo matching algorithm is used, in which a MRF model is assumed that is defined on the four-connected grid graph; the local image similarity is used for the data term and a truncated l_1 norm $f_{ij} = \max(|d_i - d_j|, d_{\max}/10)$ is used for the smoothness term. We use two types of similarity; one is the SAD-based one that is used in our method, and the other is the distance between SIFT descriptors at the corresponding points [60]. Then, the optimization of the resulting MRF models is performed using graph cuts [27]. Similarly to the above, two depth maps are computed and are differentiated to obtain scene changes. We call these procedures patch-MVS and SIFT-MVS.

5.3.2 Comparison of the results

Figure 17 shows the results for a scene. From left to right columns, the input images with a hand-marked ground truth, the results of the proposed method, PMVS2, Patch-MVS, and SIFT-MVS, respectively. For the proposed method, besides the detected changes, the change probability $p(c = 1|\cdots)$ is shown as a grey-scale image; its binarized version by a threshold $p > 0.5$ gives the result of change detection. For each of the MVS-based methods, besides the result, two estimated depths maps for the different times are shown. The detection result is their differences. Whether the scene changes or not is judged by whether the difference in its disparity is greater than a threshold. We chose 6 (disparity ranges in $[0 : 127]$) for the threshold, as it achieves the best results in the experiments. The red patches in the depth maps of PMVS2 indicate that there is no reconstructed point in the space.

Comparing the result of the proposed method with the ground truth, it is seen that the proposed method can correctly detect the scene changes, i.e., the disappearance of the debris and the digger; the shape of the digger arm is extracted very accurately. There are also some differences. The proposed method cannot detect the disappearance of the building behind the digger and of the thin layer of sands on the ground surface. The former is considered to be because the building is occluded by the digger in other viewpoints. The proposed method does not have a mechanism of explicitly dealing with occlusions but using multiple pairs of images, which will inevitably yield some errors. For the layer of sands, its structural difference might be too small for the proposed method to detect it.

The results of the MVS-based methods are all less accurate than the proposed method. As these methods differentiate the two depth maps, a slight reconstruction error in each will results in a false positive. Thus, even though their estimated depths appear to capture the scene structure mostly well, the estimated scene changes tends to be worse than the impression we have for each depth map alone.

There are in general several causes of errors in MVS-based depth estimation. For example, MVS is vulnerable to objects without textures (e.g., the ground surface in this scene). PMVS2 does not reconstruct objects that do not have reliable observations, e.g., textureless objects. As the proposed method similarly obtains depth information from image similarity, the same difficulties will have bad influence on the proposed method. However, it will be minimized by the probabilistic treatment of the depth map; taking all probabilities into account, the proposed method makes a binary decision as to whether a scene point changes or not.

We obtain precision and recall for each result using the ground truth and then calculate its F_1 score; it is 0.76, 0.59, 0.53, 0.71, in the order of Fig. 17, respectively.

Figure 18 shows results for other images. From top to bottom rows, I', the ground truths, the results of the proposed method, and those of SIFT-MVS are shown, respectively. It is seen that the proposed method produces better results for all the images. This is quantitatively confirmed by their F_1 scores which are shown in Table 1.

6. Land Surface Condition Analysis

This section presents a unified framework for robustly integrating image data taken at vastly different viewpoints to generate large-scale estimates of land surface conditions [51]. The previous sections proposed the 2D and 3D change detection methods. The method proposed in this section estimates change of debris distribution in a city based on object recognition. For recovery operation in tsunami-damaged area, it is essential to make it possible to understand debris distribution in a city.

Automated visual analysis is an effective method for understanding changes in natural phenomena over massive city-scale landscapes. However, the view-point spectrum across which image data can be acquired is extremely wide, ranging from macro-

our web site: http://www.vision.is.tohoku.ac.jp/us/download/.

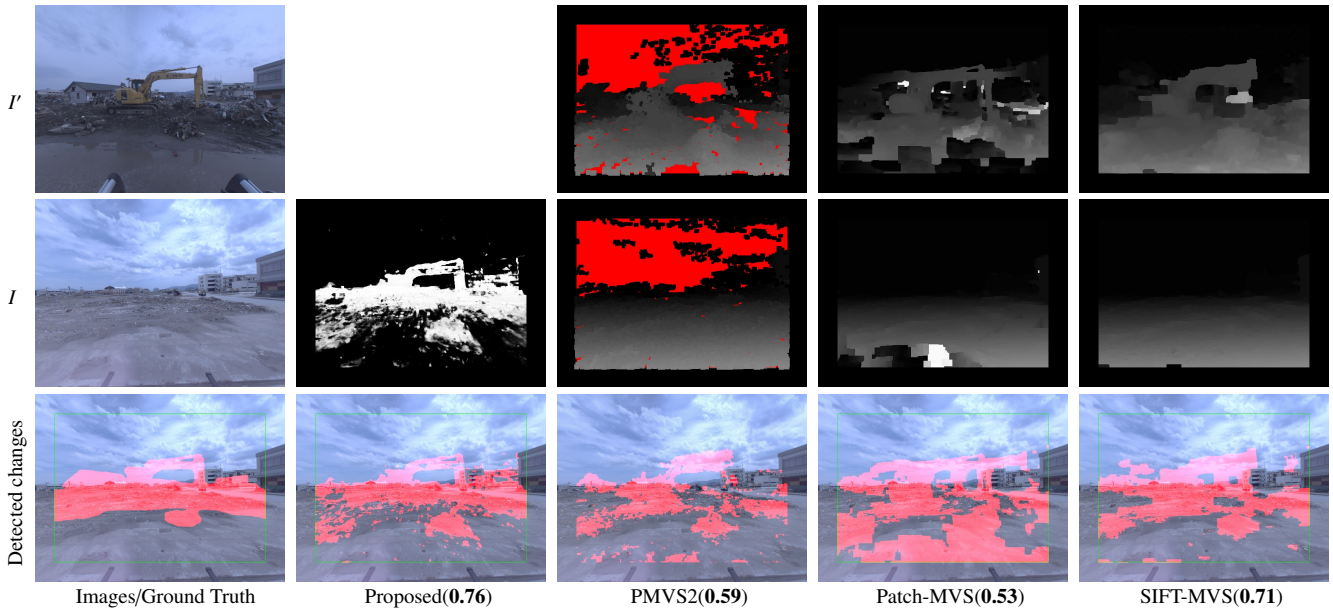


Fig. 17 Results of the proposed method and the three MVS-based ones for a scene. From left to right columns, the input images and the ground truth, the results of the proposed methods, and those of PMVS2, Patch-MVS, and SIFT-MVS, respectively. The third row shows the detected changes. The numbers in their captions are the F_1 scores representing accuracy of the detection.

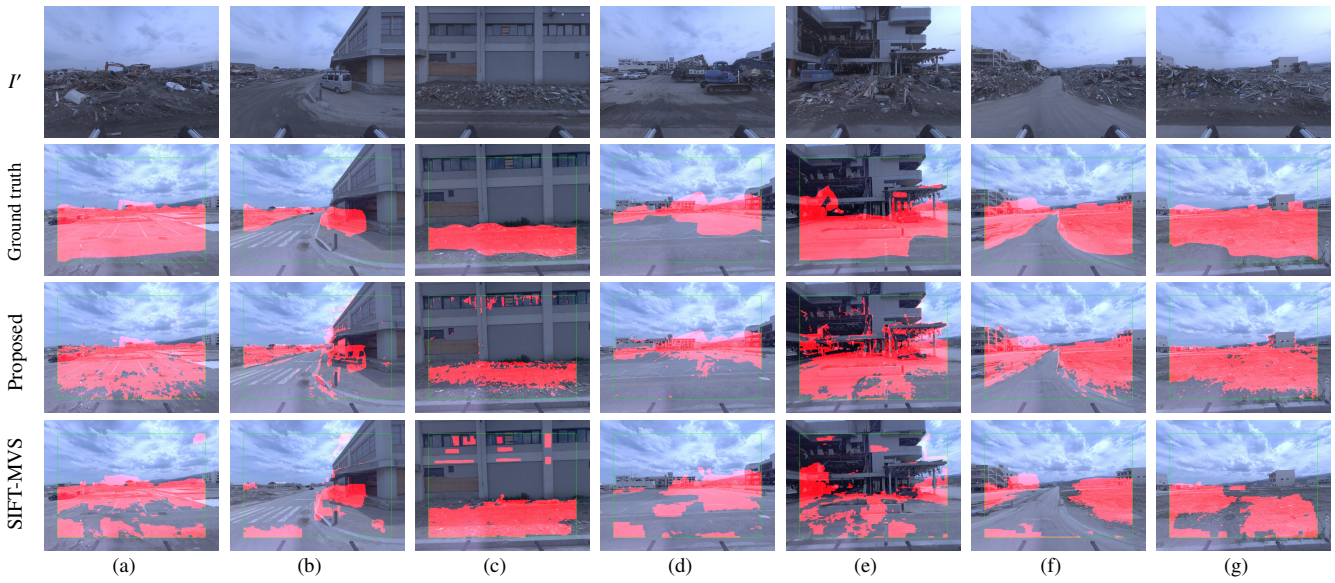


Fig. 18 Results for other images. From top to bottom rows, I', hand-marked ground truths, results of the proposed method, and those of SIFT-MVS.

Table 1 F_1 scores of the detected changes shown in Fig. 18.

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	Average
Proposed	0.88	0.67	0.77	0.85	0.82	0.91	0.92	0.83
PMVS2	0.49	0.30	0.65	0.66	0.56	0.58	0.66	0.56
Patch-MVS	0.66	0.28	0.69	0.60	0.70	0.65	0.77	0.62
SIFT-MVS	0.68	0.41	0.73	0.71	0.60	0.67	0.73	0.65

level overhead (aerial) images spanning several kilometers to micro-level front-parallel (street-view) images that might only span a few meters. To validate the proposed approach this study attempt to estimate the amount of post-tsunami damage over the entire city of Kamaishi, Japan (over 4 million square-meters). The results show that the proposed approach can efficiently integrate both micro and macro-level images, along with other forms of meta-data, to efficiently estimate city-scale phenomena. Experiments evaluate the proposed approach on two modes of land condition analysis, namely, city-scale debris and greenery estimation, to show the ability of the proposed method to generalize to a diverse set of estimation tasks.

6.1 Motivation

We address the task of estimating large-scale land surface conditions using overhead aerial (macro-level) images and street

Fig. 19 Aerial images affected by weather condition (Left: March 11, 2011, Right: March 31, 2011). The land surface might be covered by clouds and illumination conditions change drastically in aerial image.

view (micro-level) images. These two types of images are captured from orthogonal viewpoints and have different resolutions, thus conveying very different types of information that can be used in a complementary way. Moreover, their integration is necessary to make it possible to accurately understand changes in natural phenomena over massive city-scale landscapes.

Aerial images are an excellent source for collecting wide-area information of land surface conditions. However, it may come at the cost of a lower resolution (i.e., number of pixels per meter) and visiiblity may drastically change depending on the weather. For example, clouds may obscure the visibility of the land surface (Fig. 19). A more important limitation of aerial images is that they are limited to a vertical (top-down) perspective of the ground surface, such that areas occluded by a roof or highway overpass are not visible to the camera (first and second row of Fig. 20) making it difficult to estimate land conditions in covered areas.

Street-view images, on the other hand, captured from the ground-level can obtain higher resolution images of vertical structures and have better access to information about covered areas. They are also less affected by weather conditions. In the same token however, street view images are constrained to the ground plane and a single image has limited physical range. It is also labor intensive to acquire street-level images of large land surface areas (i.e., millions of square meters).

The key technical challenge is devising a method to integrate these two disparate types of image data in an effective manner, while leveraging the wide coverage capabilities of macro-level images and detailed resolution of micro-level images. The strategy proposed in the work uses macro-level imaging to learn land condition correspondences between land regions that share similar visual characteristics (e.g, mountains, streets, buildings, rivers), while micro-level images are used to acquire high resolution statistics of land conditions (e.g., the amount of debris on the ground). By combining the macro and micro level information about region correspondences and surface conditions, our proposed method generates detailed estimates of land surface conditions over the entire city.

6.2 Large-scale estimation of land surface condition

Our framework integrates aerial and street-view images to estimate land surface conditions. In this section, we explain the details of the proposed method contextualized for post-Tsunami debris detection. Although the following explanation takes debris as an example, the method is generally applicable to other types of land surface conditions. The proposed method consists of the

Fig. 20 Example aerial and street-view images. There are many cases in which aerial images and street-view images give complementary information about the land surface condition. For example, the areas covered by the building roof (the top and second row), stacked objects (the bottom row) are best viewed from the street.

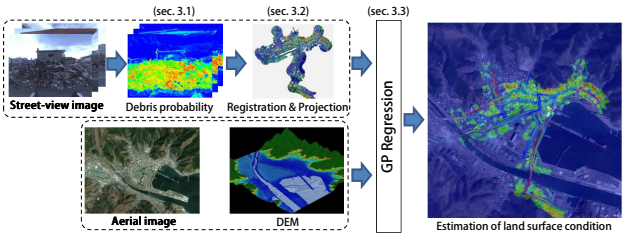


Fig. 21 Data flow diagram of city-scale estimation of land surface condition. Our approach efficiently integrates both micro (street-view) and macro-level (aerial) images along with other forms of metadata to estimate city-scale land surface condition.

following three steps;

(i) Debris detection on perspective street-view image. (sec.6.2.1)
(ii) Projection of debris probabilities on street-view images to the ground using building contours. (sec.6.2.2)
(iii) Estimation of debris over an entire city by integrating the projection result with all other data (e.g. aerial image, DEM) using a Gaussian process.(sec.6.2.3)

In the first step, the probability map of debris is calculated for each street-view image. Then, using the camera parameters for the street-view image, the probability map is projected onto the ground plane registered to a corresponding part of the aerial image. This projection method takes the existence of building walls into consideration. Finally in order to complement the estimation results obtained from street-view images, the projected probability map is integrated with the information obtained from aerial images and DEM using Gaussian process regression model.

6.2.1 Debris detection

We developed a method to calculate the probability map of debris (Fig. 22). The debris model is learned from a hand-labeled training image.The debris in the images are irregular, complicated in shape and appearance. Therefore, we exploit Geometric Context [11] as geometric feature and pixel-wise object probability [32] as an appearance feature. Geometric Context estimates the probabilities that a super-pixel belongs to seven classes. We chose four of the seven classes, ”ground plane”, ”sky”, ”porous non-planar” and ”solid non-planar”, and used the probabilities of them as debris features. The pixel-wise object probability p_{object}

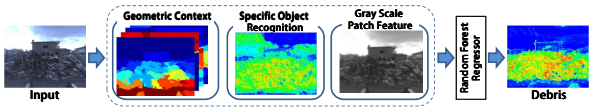
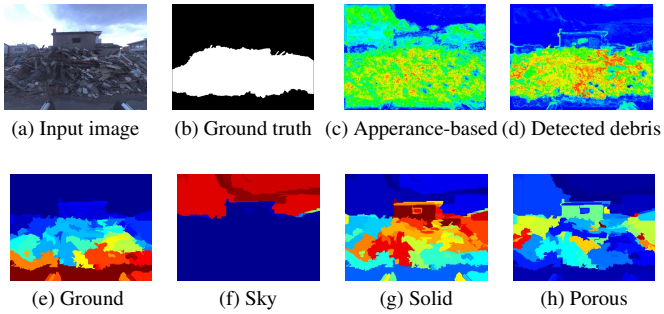


Fig. 22 Data flow diagram of debris detection. As features of debris, the probabilities of geometric context, specific object recognition and patch features are employed.



(a) Input image (b) Ground truth (c) Apperance-based (d) Detected debris

(e) Ground (f) Sky (g) Solid (h) Porous

Fig. 23 Inputs and outputs of debris detection. First rows: (a) input image. (b) hand-labeled ground truth of debris. (c) result of specific object recognition. (d) final result of debris detection. Second rows: probability of geometric context (e) ground plane, (f) sky, (g) solid non-planar, (h) porous non-planar. Color denotes probability of each class, with blue corresponding to 0 and red to 1.

is calculated using [32], Lab, HOG[9], BRIEF[5] and ORB[49]. The feature vector of debris is as follows.

$$\mathbf{x} = \left(p_{\text{ground}}, p_{\text{sky}}, p_{\text{porous}}, p_{\text{solid}}, p_{\text{object}}, m_{\text{patch}}, v_{\text{patch}}\right)^{\mathrm{T}}, \quad (1)$$

where p_{ground}, p_{sky}, p_{porous} and p_{solid} are the probabilities of "ground plane", "sky", "porous non-planar" and "solid non-planar", respectively. In addition to these probabilities, mean m_{patch} and variance v_{patch} of grayscale patch (5×5) are added to the features. Figure 23 shows an example of the datasets and detection results.

6.2.2 Projection of debris probabilities onto the ground

The debris probability explained in the previous section is the probability map on the street-view image. In order to integrate this probability map with the aerial image, the debris probability is projected onto the ground plane. Figure 24 shows the data flow diagram of projection of street-view image to the coordinate of the aerial image. The projection requires camera parameters of each street-view image. First, we performed Structure from Motion (SfM) to acquire the camera trajectories. We employ a standard SfM method [20], [34], [62] with extensions to deal with omni-directional images [61]. The estimated camera trajectories are fitted to the GPS trajectory by similarity transformations in a least squares sense.

Dividing the ground plane into a grid, we project the debris probability to the grid using projection matrix of each image. In this projection, we use the 3D models of the buildings that are generated from a 2D map of the city (Sec. 6.3.1). To be specific, the debris probability is projected to a building wall if the wall is on the projection path, and otherwise it is directly projected to the ground.

6.2.3 Integration using Gaussian Process regression

The projected debris probability map obtained up to now has no information for some areas because of occlusions or the lack of

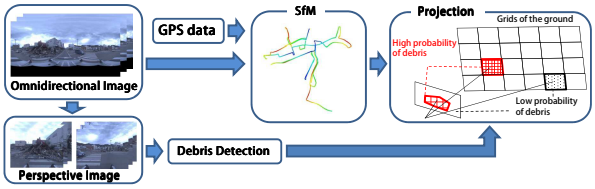


Fig. 24 Data flow diagram of the projection onto the ground plane. SfM is performed using omnidirectional street-view images. The street-view camera poses are registered to a common coordinate with aerial images and other forms of meta-data using the GPS data. After debris detection, the debris probabilities are projected to the ground plane.

street-level images. Estimating debris probability map from only an aerial image is difficult due to its low-resolution, occlusion or weather conditions. To mutually complement the street-view images and the aerial image, we used a Gaussian process regression model[46]. The main idea here is that similar geographical location tend to have similar debris probability. In the case of Tsunami-disaster, Tsunami continuously spreads from seashore to hill side, which means the damage caused by Tsunami has strong correlation with the location, especially with the elevation.

6.3 Experimental results

In order to evaluate the effectiveness of our proposed approach for estimating large-scale land conditions, we perform two experiments. Our first experiment is a comprehensive ablative analysis to examine the benefit of integrating micro and macro-level imagery for city-scale land condition estimation. In addition to color imaging, we also evaluate the contributions of two other modes of data, namely, a digital elevation map (DEM) and building occupancy maps (BOM). In our second experiment, we focus on estimating the amount of greenery and vegetation across the entire city of Kamaishi. We use the exact same approach as the debris estimation described in this paper and apply it to greenery estimation. Our results show that our approach is not limited to post-disaster analysis but can easily be applied to other modes of land condition analysis.

We created the ground truth labels used for the following evaluation by many hours of manual labeling of regions on the aerial images. Ground truth data of debris and greenery were generated by visual inspection by comparing the aerial image against the street-view images available on Google Earth. Many hours of ground truth labeling confirms that the manual inspection of large-scale land conditions is not a practical solution for real-world applications.

6.3.1 Our data

Our experiment includes two image-based input modalities and two sources of city-scale meta-data, which are described below.

Street images. We have been creating image archives of urban and residential areas damaged by Great East Japan Earthquake in 2011. The target area is 500 kilometers long along the northern-east coastal line in Japan. The images were captured every three to four months by a vehicle having an omni-directional camera (Ladybug 3 and 5 of Point Grey Research Inc.) on its roof. The image data accumulated so far amount to about 40 terabytes. The target of this experiment is the entire city of Kamaishi, Japan

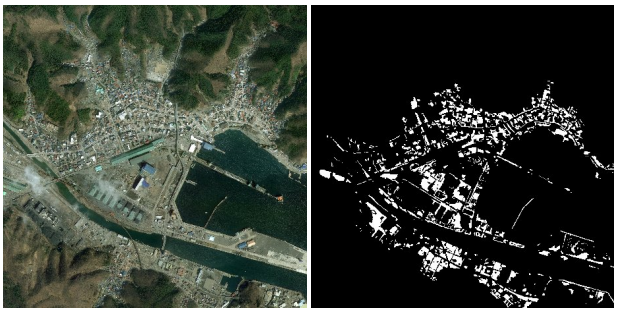


Fig. 25 Estimation target area in Kamaishi on March 31st, 2011 (left) and its hand-labeled ground truth of debris area (right). White area shows debris area.

(over 4 million square-meters). For the experiments, we chose the two image sequences captured on April 26th, 2011 (one month after the Tsunami) and August 17th, 2013 (two years and five months after the Tsunami). The debris can often be seen in the earlier images, while they tend to disappear in the later images as the recovery operation proceeds.

The street images are used for appearance-based recognition of 'stuff' [21] described in Section 6.2.1. The results of pixel-wise regression are then projected onto the ground plane as an input feature for our city-scale GP regressor.

Aerial images We downloaded aerial images from Google Earth for March 31st, 2011 and May 13th, 2012. We chose these dates to match up the timestamp of the street images.

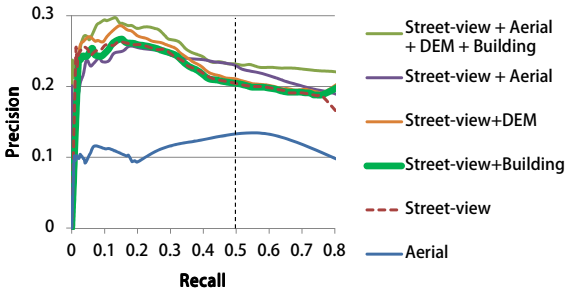
We used the aerial images for appearance-based recognition of 'stuff' categories using the same method describe in Section 6.2.1 but applied to the entire aerial image as a comparative baseline. We used the aerial images of May 13th, 2012 as the labeled training data and test on the March 31st, 2011 aerial image. Figure 25 shows an example of the hand-labeled ground truth of the debris area on the aerial images.

Digital Elevation Map (DEM). We obtained the DEM information freely available from the Geospatial Information Authority, under the Ministry of Land, Infrastructure, Transportation and Tourism in Japan. The mesh resolution of the DEM is 5×5 square-meters and contains the elevation level for each grid location. The elevation is used directly as a feature for the city-scale GP regression.

Building Occupancy Map (BOM) The BOM provides building contours. We obtained the data from Zenrin Company. The building contour data used for this experiment was made before the earthquake. We used the BOM to prevent 'stuff' from being projected onto the ground over building location.

6.3.2 Ablative Analysis

We examine the effects of each input data type on the overall performance of our proposed approach. Figure 26 shows the estimation results of the debris amounts in the entire city on April 26th, 2011 and August 17th, 2013. The lines on the aerial images are the camera trajectories. Figure 27 shows the performance of our debris detection by PR-plot and $F1$-score using different combination of input data. The results indicate that using aerial images alone yields low performance because the appearance of land conditions can change significantly over time due to changes in imaging conditions. When compared to the independent use of



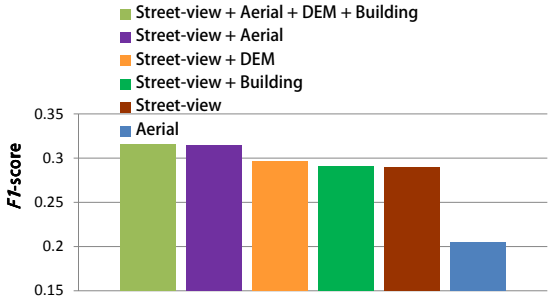


Fig. 27 Precision-recall curve of the debris area detection whose ground truth is Fig. 25. These figures show that the integration of street-view image with aerial image is efficient to estimate city-scale land surface condition.

aerial images, our results indicate that street images are more accurate for estimating city-scale debris. Furthermore, when both aerial and street images are combined we obtain better performance as the aerial information helps the city-scale GP regression to generalize to across similar looking city regions.

6.3.3 Extensions to City-Scale Vegetation Estimation

We applied our method to vegetation detection, to show how our approach can generalize to other modes of land condition estimation. Figure 28 shows an example of vegetation estimation in street-level images. The green vegetation detected in the street-view images is estimated using the same pixel-wise object recognition method [32].

Figure 29 shows the results of vegetation estimation for the entire city similar to Fig.26. By observing the vegetation heat map for the entire city, it is clear that most of the vegetation has been washed away by the Tsunami. There is also a sharp contrast between the wide spread distribution of debris and the lack of vegetation in the time period directly after the Tsunami. By 2013 however, we can see a large increase in the number of regions covered by vegetation. Our successful vegetation detection indicates that our proposed method can indeed generalize to different types of targeted estimation of city-scale land conditions.

7. Conclusion

This paper proposed the novel and practical methods for four-dimensional city modeling using vehicular imagery. The estimation target is the tsunami-damaged areas across the three prefectures whose total length is almost 400 kilometers. To estimate and visualize temporal change of the the areas, the image archive activity started one month after Great East Japan Earthquake which

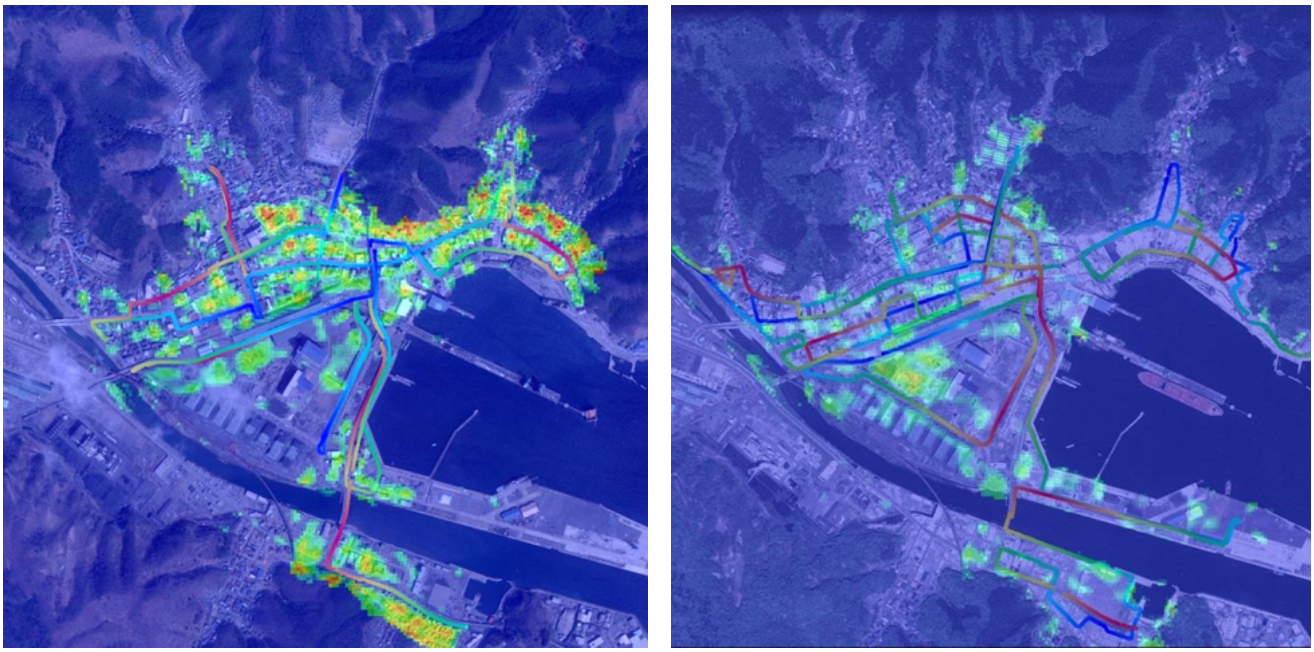


Fig. 26 City-scale **Debris** Probability in Kamaishi before the recovery operation (Left: April 26th, 2011, Right: August 17th, 2013). Color denotes probability of debris, with blue corresponding to 0 and red to 1.

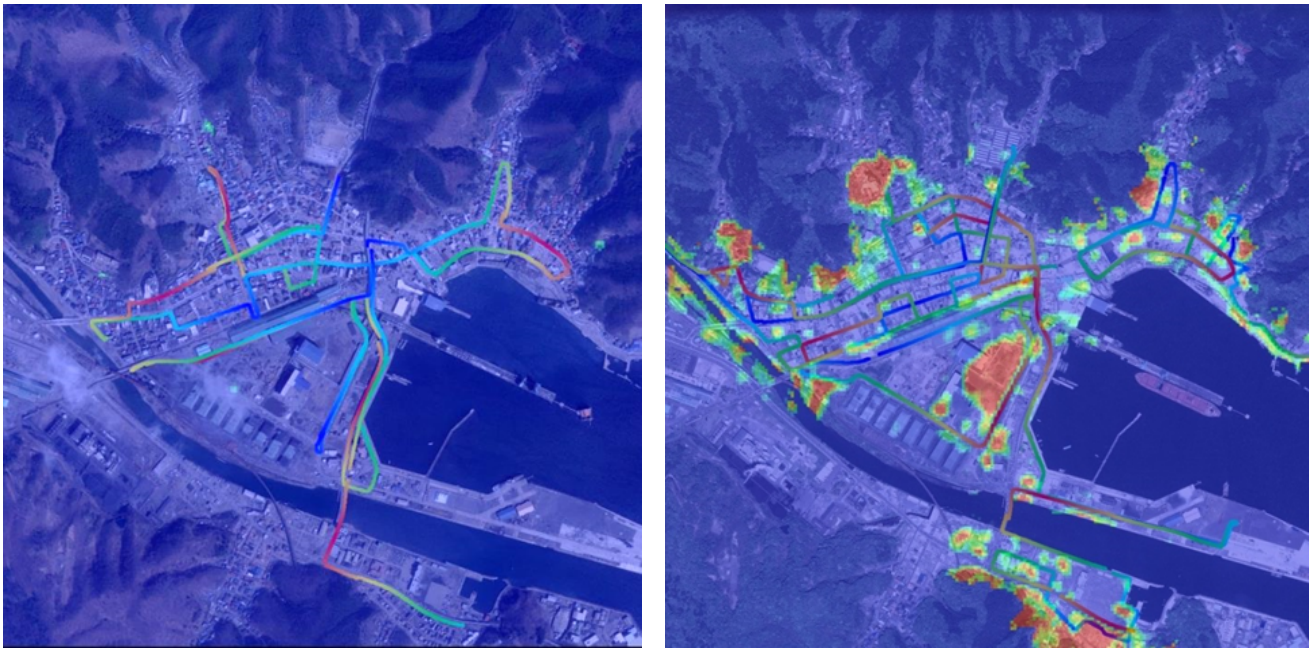


Fig. 29 City-scale **Vegetation** Probability in Kamaishi before the recovery operation (Left: April 26th, 2011, Right: August 17th, 2013). Color denotes probability of debris, with blue corresponding to 0 and red to 1.

(a) Input image (b) Ground truth (c) Detected Vegetation

Fig. 28 Green vegetation detection. (a) input image. (b) hand-labeled ground truth of green vegetation. (c) probability of green vegetation. Color denotes probability of green vegetation, with blue corresponding to 0 and red to 1.

caused the giant Tsunami. The Tsunami gave serious damages to the Pacific coast area of the Tohoku. The images periodically recorded the the scenes of the tsunami-damaged areas.

From the periodic images, this paper visualized the tsunami-damage and recovery of the tsunami-damaged area. First, the 2D change detection method using grid feature roughly but quickly estimates scene change of entire areas. Next, the structural change detection method estimates more accurate scene change even if there is ambiguity in estimated scene depth. Finally, the method of land surface condition analysis estimates city-scale temporal change integrating aerial and vehicular imagery.

The 2D method detects scene change using grid feature from an image pair without 3D model and pixel-level registration. The experimental results show the effectiveness of the proposed method integrating high discrimination of convolutional neural network (CNN) feature with accurate segmentation of superpixel in 2D change detection. As a by-product, the method can reduce the computational time.

The structural change detection method detects temporal changes of the three dimensional structure of an outdoor scene from its multi-view images captured at two separate times. The method estimates scene structures probabilistically, not deterministically to maximize the accuracy of change detection. The proposed method is compared with the methods that use multi-view stereo (MVS) to reconstruct the scene structures of the two time points and then differentiate them to detect changes. The experimental results show that the proposed method outperforms such MVS-based methods. Unlike MVS-based methods, the proposed method can estimate accurate shape of the scene change (e.g. debris) because the proposed method utilizes no prior on the smoothness of scene structure.

The method of land surface condition analysis is a unified framework for robustly integrating image data taken at vastly different viewpoints to generate large-scale estimates of land surface conditions. The method uses macro-level imaging to learn land condition correspondences between land regions that share similar visual characteristics, while micro-level images are used to acquire high resolution statistics of land conditions. The experimental results show that the proposed approach can effectively integrate both macro (aerial) and micro-level (vehicular) images, along with other forms of meta-data, to estimate city-scale phenomena. Furthermore, the proposed method can be successfully applied to vegetation estimation. The results indicate the method can generalize well to many kinds of applications to estimate city-scale phenomena by replacing the detector target (e.g. human flow, real-estate and dirt quality).

This paper achieved the objective of developing the methods for 4D city modeling in tsunami-damaged area using vehicular imagery. As mentioned in section 1, to estimate temporal change of regional-scale area using vehicular imagery, there are three challenges to overcome as follows, (i) limited camera viewpoint, (ii) limited physical range, (iii) large computation. The 3D change detection method makes it possible to detect structural change even if there is depth ambiguity due to the limited camera viewpoint. The method of land surface condition analysis integrates aerial and vehicular imagery and estimates change of debris distribution for entire city. Furthermore, the 2D change detection method can reduce the computational time and makes it possible to process the entire tsunami-damaged areas with a single workstation.

For future work, the three methods mentioned above can be integrated into a system which estimates temporal changes of vastly wide area, for example, the entire tsunami-damaged areas of the Tohoku. It is possible for all the methods to process multiple areas in parallel. If multiple computers are available, the temporal change of the entire tsunami-damaged areas can be estimated in a day or a few days.

If the number of sensors increases in the future (e.g. came mounted on self-driving car), scene images of cities will be available in real-time. The real-time sensor networks can generate real-time 3D map [30] and apply statistical analysis. The proposed 4D modeling approach is fast enough to be applied to such on-line sensory information. Combined with the real-time big data, the proposed method can extend to real-time monitoring of the city.

References

[1] Agarwal, S., Snavely, N., Simon, I., Seitz, S. M. and Szeliski, R.: Building Rome in a day, *ICCV*, pp. 72–79 (2009).
[2] Bay, H., Tuytelaars, T. and Van Gool, L.: Surf: Speeded Up Robust Features, *ECCV*, Springer, pp. 404–417 (2006).
[3] Berni, J. A. J., Member, S., Zarco-tejada, P. J., Suárez, L. and Fereres, E.: Thermal and Narrowband Multispectral Remote Sensing for Vegetation Monitoring From an Unmanned Aerial Vehicle, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 47, No. 3, pp. 722–738 (2009).
[4] Cabezas, R., Freifeld, O., Rosman, G. and Fisher III, J. W.: Aerial Reconstructions via Probabilistic Data Fusion, *IEEE Computer Vision and Pattern Recognition Conference on Computer Vision* (2014).
[5] Calonder, M., Lepetit, V., Strecha, C. and Fua, P.: BRIEF : Binary Robust Independent Elementary Features , *ECCV* (2010).
[6] Campbell, J. B. and Wynne, R. H.: *Introduction to Remote Sensing (5th edition)*, Guilford Press (2011).
[7] Crandall, D., Owens, A., Snavely, N. and Huttenlocher, D.: Discrete-Continuous Optimization for Large-Scale Structure from Motion, *CVPR*, pp. 3001–3008 (2011).
[8] Crispell, D., Mundy, J. and Taubin, G.: A Variable-Resolution Probabilistic Three-Dimensional Model for Change Detection, *Geoscience and Remote Sensing*, Vol. 50, No. 2, pp. 489–500 (2012).
[9] Dalal, N. and Triggs, B.: Histograms of Oriented Gradients for Human Detection, *CVPR*, pp. 886–893 (2005).
[10] Delenne, C., Durrieu, S., Rabatel, G. and Deshayes, M.: From pixel to vine parcel: A complete methodology for vineyard delineation and characterization using remote-sensing data, *Computers and Electronics in Agriculture*, Vol. 70, No. 1, pp. 78–83 (2010).
[11] Derek Hoiem, Alexei A. Efros and Martial Hebert: Geometric context from a single image, *ICCV*, pp. 654–661 (2005).
[12] Fischler, M. A. and Bolles, R. C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, *Communications of the ACM*, Vol. 24, No. 6, pp. 381–395 (1981).
[13] Fischler, M. A. and Bolles, R. C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, *Communications of the ACM*, Vol. 24, No. 6, pp. 381–395 (1981).
[14] Frahm, J.-M., Fite-Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.-H., Dunn, E., Clipp, B., Lazebnik, S. et al.: Building Rome on a cloudless day, *Computer Vision–ECCV 2010*, Springer, pp. 368–381 (2010).
[15] Furukawa, Y. and Ponce, J.: Accurate, Dense, and Robust Multi-View Stereopsis, *PAMI*, Vol. 32, No. 8, pp. 1362–1376 (2010).
[16] Gong, P., Pu, R. and Chen, J.: Mapping Ecological Land Systems and Classification Uncertainties from Digital Elevation and Forest-Cover Data Using Neural Networks, *Photogrammetric Engineering & Remote Sensing*, Vol. 62, No. 11, pp. 1249–1260 (1996).
[17] Haala, N. and Kada, M.: An update on automatic 3D building reconstruction, *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 65, No. 6, pp. 570–580 (2010).
[18] Hall, A., Louis, J. and Lamb, D.: Characterising and mapping vineyard canopy using high-spatial-resolution aerial multispectral images, *Computers & Geosciences*, Vol. 29, No. 7, pp. 813–822 (2003).
[19] Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O. and Townshend, J. R. G.: High-Resolution Global Maps of 21st-Century Forest Cover Change, *Science*, Vol. 342, pp. 850–853 (2013).
[20] Hartley, R. and Zisserman, A.: *Multiple View Geometry in Computer Vision Second Edition*, Cambridge University Press (2004).
[21] Heitz, G. and Koller, D.: Learning spatial context: Using stuff to find things, *Computer Vision–ECCV 2008*, Springer, pp. 30–43 (2008).
[22] Herold, M., Liu, X. and Clarke, K. C.: Spatial Metrics and Image Texture for Mapping Urban Land Use, *Photogrammetric Engineering & Remote Sensing*, No. 9, pp. 991–1001 (2003).

[23] Hu, J., You, S. and Neumann, U.: Approaches to large-scale urban modeling, *Computer Graphics and Applications, IEEE*, Vol. 23, No. 6, pp. 62–69 (2003).
[24] Huertas, A. and Nevatia, R.: Detecting Changes in Aerial Views of Man-Made Structures, *ICCV*, pp. 73–80 (1998).
[25] Ibrahim Eden, D. C.: Using 3D Line Segments for Robust and Efficient Change Detection from Multiple Noisy Images, *ECCV*, pp. 172–185 (2008).
[26] Kaminsky, R., Snavely, N., Seitz, S. and Szeliski, R.: Alignment of 3D Point Clouds to Overhead Images, *CVPR Workshops*, pp. 63–70 (2009).
[27] Kolmogorov, V. and Zabih, R.: What energy functions can be minimized via graph cuts?, *PAMI*, Vol. 26, No. 2, pp. 147–59 (2004).
[28] Lafarge, F. and Mallet, C.: Building large urban environments from unstructured point data, *Computer Vision (ICCV), 2011 IEEE International Conference on*, IEEE, pp. 1068–1075 (2011).
[29] Lafarge, F. and Mallet, C.: Creating large-scale city models from 3D-point clouds: a robust approach with hybrid representation, *International journal of computer vision*, Vol. 99, No. 1, pp. 69–85 (2012).
[30] Lee, K.-H., Hwang, J.-N., Okapal, G. and Pitton, J.: Driving recorder based on-road pedestrian tracking using visual SLAM and Constrained Multiple-Kernel, *International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, pp. 2629–2635 (2014).
[31] Li, F., Jacksona, T. J., Kustasa, W. P., Schmuggea, T. J., Frenchb, A. N., Cosha, M. H. and Bindlish, R.: Deriving land surface temperature from Landsat 5 and 7 during SMEX02/SMACEX, *Remote Sensing of Environment*, Vol. 92, No. 4, pp. 521–534 (2004).
[32] Li, Cheng and Kitani, Kris M.: Pixel-level Hand Detection in Ego-Centric Videos, *CVPR*, pp. 3570–3577 (2013).
[33] Lin, C. and Nevatia, R.: Building Detection and Description from a Single Intensity Image, *Computer Vision and Image Understanding*, Vol. 72, No. 2, pp. 101–121 (1998).
[34] Lowe, D. G.: Distinctive Image Features from Scale-Invariant Keypoints, *IJCV*, Vol. 60, No. 2, pp. 91–110 (2004).
[35] Lu, D., Hetrick, S. and Moran, E.: Impervious surface mapping with QuickBird imagery, *International journal of remote sensing*, Vol. 32, No. 9, pp. 2519–2533 (2011).
[36] Martinez, J. and Letoan, T.: Mapping of flood dynamics and spatial distribution of vegetation in the Amazon floodplain using multitemporal SAR data, *Remote Sensing of Environment*, Vol. 108, No. 3, pp. 209–223 (2007).
[37] Musialski, P., Wonka, P., Aliaga, D. G., Wimmer, M., van Gool, L. and Purgathofer, W.: A Survey of Urban Reconstruction, *Computer Graphics Forum*, Vol. 32, No. 6, pp. 146–177 (online), DOI: 10.1111/cgf.12077 (2013).
[38] Nistér, D.: An efficient solution to the five-point relative pose problem, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 26, No. 6, pp. 756–770 (2004).
[39] Pollard, T. and Mundy, J. L.: Change Detection in a 3-d World, *CVPR*, pp. 1–6 (2007).
[40] Pollefeys, M., Nistér, D., Frahm, J.-M., Akbarzadeh, A., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Kim, S.-J., Merrell, P., Salmi, C., Sinha, S., Talton, B., Wang, L., Yang, Q., Stewénius, H., Yang, R., Welch, G. and Towles, H.: Detailed Real-Time Urban 3D Reconstruction from Video, *IJCV*, Vol. 78, No. 2-3, pp. 143–167 (2008).
[41] Poullis, C. and You, S.: Automatic creation of massive virtual cities, *Virtual Reality Conference, 2009. VR 2009. IEEE*, IEEE, pp. 199–202 (2009).
[42] Poullis, C. and You, S.: Automatic reconstruction of cities from remote sensor data, *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, IEEE, pp. 2775–2782 (2009).
[43] Poullis, C. and You, S.: Photorealistic large-scale urban city model reconstruction, *Visualization and Computer Graphics, IEEE Transactions on*, Vol. 15, No. 4, pp. 654–669 (2009).
[44] Poullis, C. and You, S.: 3d reconstruction of urban areas, *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2011 International Conference on*, IEEE, pp. 33–40 (2011).
[45] Radke, R. J., Andra, S., Al-Kofahi, O. and Roysam, B.: Image Change Detection Algorithms: A Systematic Survey, *Transactions on Image Processing*, Vol. 14, No. 3, pp. 294–307 (2005).
[46] Rasmussen, C. E. and Williams, C. K. I.: *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press (2005).
[47] Rottensteiner, F., Sohn, G., Gerke, M. and Wegner, J. D.: ISPRS test project on urban classification and 3D building reconstruction, *Commission III-Photogrammetric Computer Vision and Image Analysis, Working Group III/4-3D Scene Analysis*, pp. 1–17 (2013).
[48] Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S. and Breitkopf, U.: The ISPRS benchmark on urban object classification and 3D building reconstruction, *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences I-3*, pp. 293–298 (2012).
[49] Rublee, E., Rabaud, V., Konolige, K. and Bradski, G.: ORB : an efficient alternative to SIFT or SURF, *ICCV*, pp. 2564–2571 (2011).
[50] Sakurada, K., Okatani, T. and Deguchi, K.: Detecting Changes in 3D Structure of a Scene from Multi-view Images Captured by a Vehicle-Mounted Camera, *CVPR*, pp. 137–144 (2013).
[51] Sakurada, K., Okatani, T. and Kitani, K. M.: Massive City-scale Surface Condition Analysis using Ground and Aerial Imagery, *ACCV* (2014).
[52] Schindler, G. and Dellaert, F.: Probabilistic temporal inference on reconstructed 3D scenes, *CVPR*, pp. 1410–1417 (2010).
[53] Schowengerdt, R. A.: *Remote Sensing: Models and Methods for Image Processing* (2006).
[54] Snavely, N., Seitz, S. M. and Szeliski, R.: Modeling the World from Internet Photo Collections, *IJCV*, Vol. 80, No. 2, pp. 189–210 (2007).
[55] Solberg, A. H. S.: Contextual Data Fusion Applied to Forest Map Revision, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 37, No. 3, pp. 1234–1243 (1999).
[56] Strecha, C., Pylvanainen, T. and Fua, P.: Dynamic and Scalable Large Scale Image Reconstruction, *CVPR*, pp. 406–413 (2010).
[57] Suveg, I. and Vosselman, G.: Reconstruction of 3D building models from aerial images and maps, *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 58, No. 3-4, pp. 202–224 (2004).
[58] Taneja, A., Ballan, L. and Pollefeys, M.: Image based detection of geometric changes in urban environments, *ICCV*, pp. 2336–2343 (2011).
[59] Taneja, A., Ballan, L. and Pollefeys, M.: City-Scale Change Detection in Cadastral 3D Models Using Images, *CVPR*, pp. 113–120 (2013).
[60] Tola, E., Lepetit, V. and Fua, P.: DAISY: An Efficient Dense Descriptor Applied to Wide-Baseline Stereo, *PAMI*, Vol. 32, No. 5, pp. 815–830 (2010).
[61] Torii, A., Havlena, M. and Pajdla, T.: From Google Street View to 3D City Models, *ICCV Workshops*, pp. 2188–2195 (2009).
[62] Triggs, B., McLauchlan, P., Hartley, R. and Fitzgibbon, A.: Bundle Adjustment - Modern Synthesis, *ICCV*, pp. 298–372 (1999).
[63] van der Sande, C., de Jong, S. and a.P.J. de Roo: A segmentation and classification approach of IKONOS-2 imagery for land cover mapping to assist flood risk and flood damage assessment, *International Journal of Applied Earth Observation and Geoinformation*, Vol. 4, No. 3, pp. 217–229 (2003).
[64] Weng, Q.: *Remote Sensing of Impervious Surfaces*, CRC Press (2010).
[65] Weng, Q., Lu, D. and Schubring, J.: Estimation of land surface temperature–vegetation abundance relationship for urban heat island studies, *Remote Sensing of Environment*, Vol. 89, No. 4, pp. 467–483 (2004).
[66] Zebedin, L., Klaus, A., Gruber-Geymayer, B. and Karner, K.: Towards 3D map generation from digital aerial images, *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 60, No. 6, pp. 413–427 (2006).
[67] Zhang, C., Wang, L. and Yang, R.: Semantic Segmentation of Urban Scenes Using Dense Depth Maps, *ECCV*, pp. 708–721 (2010).
[68] Zhang, G., Jia, J., Xiong, W., Wong, T.-T., Heng, P.-A. and Bao, H.: Moving Object Extraction with a Hand-held Camera, *ICCV*, pp. 1–8 (2007).
[69] Zhou, Q.-Y. and Neumann, U.: A streaming framework for seamless building reconstruction from large-scale aerial lidar data, *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, IEEE, pp. 2759–2766 (2009).
[70] Zhou, Q.-Y. and Neumann, U.: 2.5 D building modeling with topology control, *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, IEEE, pp. 2489–2496 (2011).