Automated Content-Based Filtering for Enhanced Vision-based Documentation in Construction toward Exploiting Big Visual Data from Drones

Youngjib Ham[1] and Mirsalar Kamari[2]

Abstract

In recent years, emerging mobile devices and camera-equipped platforms have offered a great convenience to visually capture and constantly document the as-is status of construction sites. In this regard, visual data are regularly collected in the form of numerous photos or lengthy videos. However, massive amounts of visual data that are being collected from jobsites (e.g., data collection on daily or weekly bases by Unmanned Aerial Vehicles, UAVs) has provoked visual data overload as an inevitable problem to face. To address such data overload issue in the construction domain, this paper aims at proposing a new method to automatically retrieve photo-worthy frames containing construction-related contents that are scattered in collected video footages or consecutive images. In the proposed method, the presence of objects of interest (i.e., construction-related contents) in given image frames are recognized by the semantic segmentation, and then scores of the image frames are computed based on the spatial composition of the identified objects. To improve the filtering performance, high-score image frames are further analyzed to estimate their likelihood to be intentionally taken. Case studies in two construction sites have revealed that the accuracy of the proposed method is close-to-human judgment in filtering visual

[1] Assistant Professor, Department of Construction Science, Texas A&M University, 3137 TAMU, College Station, TX 77843; Email: yham@tamu.edu

[2] Ph.D. Student, Department of Construction Science, Texas A&M University, 3137 TAMU, College Station, TX 77843; Email: kamari@tamu.edu

data to retrieve photo-worthy image frames containing construction-related contents. The performance metrics demonstrate around 91% of accuracy in the semantic segmentation, and we observed enhanced human-like judgment in filtering construction visual data comparing to prior works. It is expected that the proposed automated method enables practitioners to assess the as-is status of construction sites efficiently through selective visual data, thereby facilitating data-driven decision making at the right time.

Keywords: Visual Sensing; Visual Data Filtering; UAV; Construction Monitoring

1. INTRODUCTION

With the increasing availability of camera-equipped devices such as smartphones, head-mounted cameras, and Unmanned Aerial Vehicles (UAVs), large numbers of high-quality images or video footages are constantly being collected to document the as-is status of construction jobsites. For instance, it was observed that for ~ 750,000 square feet of a construction jobsite, more than 400,000 images are generally documented during the lifetime of the project [1]. In general, visual data on construction sites can be collected in two ways: 1) video recording; and 2) point-and-shoot [1]. The video recording refers to recording videos or taking pictures with a constant interval using camcorders, closed circuit televisions (CCTV), or UAVs. In this way, continuing changes in construction sites can be recorded. However, the volume of visual data would be rapidly increasing proportional to its recording time, therefore it is not trivial to process a large amount of visual data for generating useful information. The point-and-shoot represents taking pictures while a camera is purposefully pointed at an object or a region of interest. Since site personnel take pictures with the intention of documenting a certain situation at a construction site, preprocessing to select

meaningful image frame would be unnecessary in general.

Typically, the amount of image data could be significantly increased when recording high-definition videos on a regular basis, involving laborious challenges for practitioners to overcome. It was observed that the average size of visual data per time instance of data collection for three construction sites amounts to 3.8 Gigabytes [1]. Furthermore, they reported that the collected visual data size could grow up to 43 Gigabytes after 3D dense point cloud reconstruction (proposed by [2]) using the collected images. Assuming weekly visual data collection for 144 weeks of construction, the size of collected visual data for a single construction project would approximately reach to 6.2 Terabytes, and the size of visual data may even reach to the scale of Petabytes, for a company managing multiple projects [1]. The large portion of such visual data captured via UAVs or head-mounted cameras may still include non-construction related contents or has been unintentionally taken during a walk-through (head-mounted cameras) or flying (UAVs). In this context, construction companies are struggling to reduce the size of recorded visual data for identifying handful of informative image frames that can better describe the as-is status of construction jobsites. Due to its huge amounts, handling large-scale visual data collected for a long period of time is not a trivial task. Filtering visual data to select meaningful frames is typically a time-consuming and labor-intensive process, which requires several hours of efforts [3,4]. In this sense, automated methods for retrieving relevant contents without human interventions has been considered as an essential task [5]. The authors have carried out preliminary experiments to estimate the time spent on manually identifying informative image frames in long-sequence video footages obtained from construction jobsites. We observed that around 1h 15m was spent by an experienced individual to select 234 informative image frames from a video footage with the size of 4.5 Gigabytes (total length of 2h 15m). This result implies that although a small portion over a

large-scale visual dataset would be sufficient to understand jobsite context, it is challenging and laborious to select meaningful image frames manually, even for a single video. The required time and effort in visual data filtering could be significantly increased for layperson or depending on the size of visual data.

To address this issue, we propose a novel visual data filtering method to automatically retrieve photo-worthy image frames containing construction-related contents from a large-scale visual dataset from drones. The proposed method consists of two stages. At the first filtering stage, the presence of objects of interest in images is identified by the semantic segmentation. This information is then used to score each image frame based on the spatial composition of the identified objects. The second filtering stage is to select intentionally-taken images among high-score images based on their compositional information. Case studies were conducted to evaluate the effectiveness of the proposed method, and the experimental results are compared with previous methods. The contribution of this study can be summarized as two-fold: 1) proposing the effective method of retrieving construction-related images based on the presence of construction-related objects and their spatial composition through the semantic segmentation with the small number of training images; and 2) demonstrating the way to select intentionally-taken photos containing construction-related contents based on their compositional similarities with around 130,000 intentionally taken images in the SUN dataset. The significance of this work is to save time and efforts for practitioners to obtain concise and yet useful images representing the as-is status of construction jobsites in an automated manner, which enables the construction industry to reduce visual data size efficiently. By being able to focus on selective visual data from jobsites, practitioners can spend less time on browsing large amounts of visual data, rather spend more time on visual analytics of various construction performance metrics using the selected dataset.

2. RELATED WORKS

Analyzing large-scale visual data has been widely studied in the construction domain. For instance, [1] demonstrated the potential of big visual data for monitoring construction performance. To leverage the potential for an efficient visual data management, automated retrieval of the most relevant contents from large-scale video footages is essential [6,7]. One of the earliest attempts was to select key image frames to generate 3D point clouds using the Structure-from-Motion (SfM) [8]. Since all image frames in a raw dataset are not necessarily used for generating 3D point cloud models, a significant amount of time required for the 3D reconstruction could be saved with the reduced numbers of images while maintaining the quality of the 3D reconstruction of built environments. Despite the benefit of such early studies providing an encouraging proof-of-concept, there have been limited attempts to research on key frame selection that is beneficial to construction practitioners for enhancing situational awareness in jobsites. As the computational efficiency of visual data filtering would be a burden due to the increasing volume of visual data from jobsites, filtering methods should be carefully designed and implemented [5].

Visual data filtering has recently attracted much attention in the computer vision domain. The types of visual data filtering techniques can be divided into three-fold [9]: 1) keyframe selection, 2) video skimming, and 3) time-lapse. Previous studies on the keyframe selection have mainly focused on selecting key video frames to represent the core content of videos [10-13]. The main application of this technique is to automatically choose representative frames describing overall video contents or to insert advertisements in between most important frames. The video skimming summarizes a video to convey its main idea in a shortened video clip [14,15]. The time-lapse (or hyper-lapse [16]) technique samples highly distinctive video frames and produces a

shortened video while retaining the smooth transitions of camera perspectives to enable users to have an immersive viewing experience. The fundamental challenge of the visual data filtering techniques is to establish criteria to select certain visual contents considering visual characteristics of videos, filtering context, and computational resources. In the construction domain, the keyframe selection has the potential for minimizing redundant contents out of large-scale visual data, while preserving valuable visual contents. Nonetheless, previous keyframe selection methods have not been tested on construction image dataset that has distinct visual characteristics, thus, their applicability for construction monitoring remains largely unknown. Particularly, as the increased availability of UAVs has recently produced a large amount of visual data from construction sites, there is an increasing need to extract the important visual contents of jobsites in an automated manner.

Generally, the keyframe selection methods can be divided into two categories considering the viewpoints of video recording: third-person and first-person viewpoints. Videos recorded by the third-person viewpoint are intentionally taken in general, and the goal of the key frame selection for such videos is generally to select video frames containing objects or scenes of interest [17,18]. Contrarily, videos recorded by the first-person viewpoint through head-mounted or wearable cameras are most likely to involve numerous frames that are not intentionally taken; therefore, the keyframe selection purpose for this category is generally to eliminate such unintentionally taken video frames from original videos. To do that, prior works such as [4] leveraged great numbers of intentionally taken photos as the benchmark dataset and then computed the compositional similarity between the benchmark dataset and testing images. Some studies utilized large-scale training image datasets collected from the Amazon Mechanical Turk [19,20] or automatic web mining [21,22] for visual data filtering. Nonetheless, in the construction domain,

such large-scale training image datasets are relatively limited and hard to be collected as reported in [1,23]. Such difficulty would hinder the direct implementation of existing filtering methods for the construction domain. In addition, distinct visual characteristic of construction images needs to be considered for effective visual data filtering in the construction domain.

3. OVERVIEW OF THE METHOD

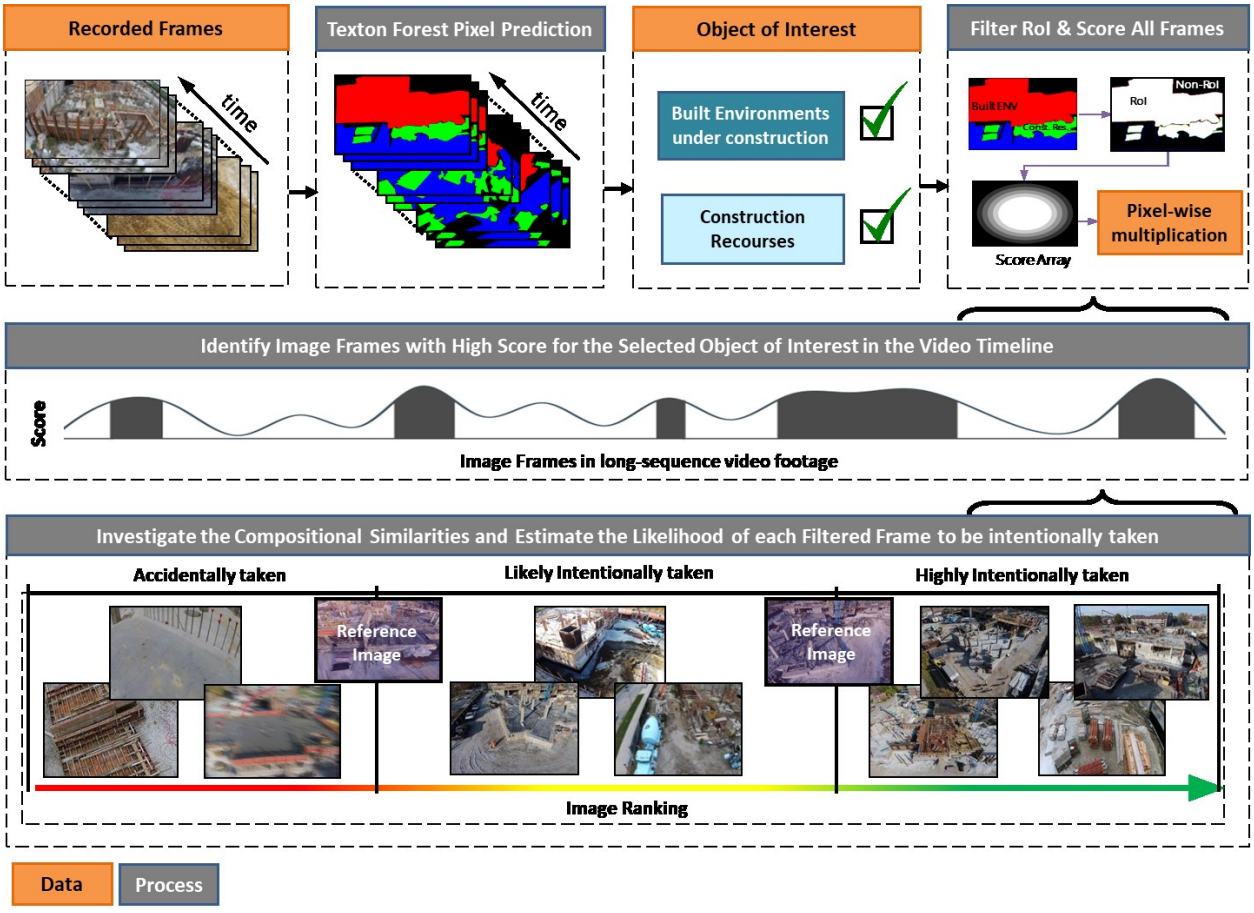


Figure 1: An overview of the proposed method to filer visual data captured from jobsites.

Figure 1 outlines the overview of the proposed method to select key frames from construction visual data. This paper aims to select key frames from large-scale aerial video footages captured via UAVs around jobsites. We define the term “key frames” as “photo-worthy frames” that can

148 provide valuable information regarding situational awareness in construction sites. In this study,
149 photo-worthy image frames should involve objects of interest (e.g., built environments under
150 construction or construction resources such as equipment or materials) appeared to be large in the
151 middle of the frames, in order to increase the chance of illustrating the as-is status of construction.

152 The proposed filtering method consists of two modules. The first filtering process retrieves
153 image frames containing objects of interest (i.e., construction-related contents) through the
154 semantic segmentation and then assigns a score to each frame based on the position of the
155 identified objects of interest in the frame; this process tightens the search space of large-scale
156 visual data, thereby increasing the chance to obtain construction-related photo-worthy images at
157 the second filtering process. The second filtering process investigates the quality of the retrieved
158 images in terms of their compositional similarities comparing with the SUN dataset, in order to
159 select intentionally-taken photos.

160 To detect objects of interest in the first filtering process, we semantically segment an image
161 into pixels by their categories through semantic texton forests building on [24] in which low level
162 texton features and decision trees are employed. This supervised method rapidly segments pixels
163 by their category with high accuracy [25], based on fewer numbers of training images compared
164 to deep learning approaches such as the deep convolutional encoder-decoder [26]. Building upon
165 the outcome of the semantic segmentation, we assign a score for each image frame based on the
166 position of objects of interest. The closer the specific objects of interest are to the center of an
167 image; the higher score is assigned. Then, we retrieve selective images based on a threshold for
168 the scores. The thresholding criterion is discussed in Section 5. In the second filtering process, the
169 compositional similarities of the retrieved images with the SUN dataset are calculated building on
170 five different visual features—Dense-SIFT, HOG, GIST, SSIM, and motion blur—to select

intentionally-taken (i.e., snap point) images. To efficiently compare the feature spaces between the two domains (i.e., the retrieved images and the SUN dataset images), the Principal Component Analysis (PCA) [27] is employed to reduce the number of the feature dimensions. Next, the similarities across the feature spaces are computed for each retrieved frame to select final images. In the following sections, the algorithmic procedure of the proposed method is described in detail, and the experimental results are then discussed.

4. RETRIEVAL OF OBJECTS OF INTEREST THROUGH THE SEMANTIC SEGMENTATION

Conventional approaches to address image segmentation problems typically involve the extraction of features from a training dataset to form a bag of visual words and train a classifier based on the frequency of visual words in each image [28-30]. Other approaches involve the extraction of a bank of descriptors and features from a training dataset and then train discriminative classifiers based on their response [31-33]. Despite the benefit, they are computationally expensive in extracting image descriptors and features for both training and testing domains. To initially filter given visual data, we build upon the semantic texton forest to segment each image into numerous regions with labels based on low-level texton features. The semantic texton forest is a low-level visual feature, which enables computationally efficient semantic segmentation than algorithms that need the expensive computation for constructing filter-bank responses or local descriptors [24].

4.1 Building Decision Trees

The semantic segmentation is the process of separating each image region by its category. Specifically, the semantic segmentation assigns each pixel **p** to a category **c** based on the visual

194 information of surrounding pixels. Through the first level decision tree, the visual information of
195 image pixels is extracted in the form of semantic textons. They are used to create the visual feature
196 distribution of each pixel for determining a pixel category. The second level decision tree is used
197 to explore the layout and context of the semantic regions for better segmentation. Figure 2
198 represents examples of decision trees for different pixel categories such as buildings under
199 construction and soil grounds. The first level decision tree contains n nodes and l leaf nodes. The
200 term leaf node corresponds to the last binary decision made in the branch of a tree. There is a
201 learned category distribution $P(c|n)$ associated with each leaf node. Based on the class distributions
202 over the leaf nodes of all trees $L(p) = (l_1, l_2, l_3, \ldots, l_T)$, the final decision on the first decision tree is
203 made; this decision process is formulated as Eq. 1. The outcome of the first level decision tree is
204 a rough estimation of a pixel category. At the second decision tree, textural relations are studied
205 based on the semantic texton histogram and the bag of semantic texton region priors to generate
206 coherent segmentation results.
207

$$P\big(c\big|L(p)\big) = \sum_{t=1}^{T} P(c|l_t)P(t) \tag{1}$$

208

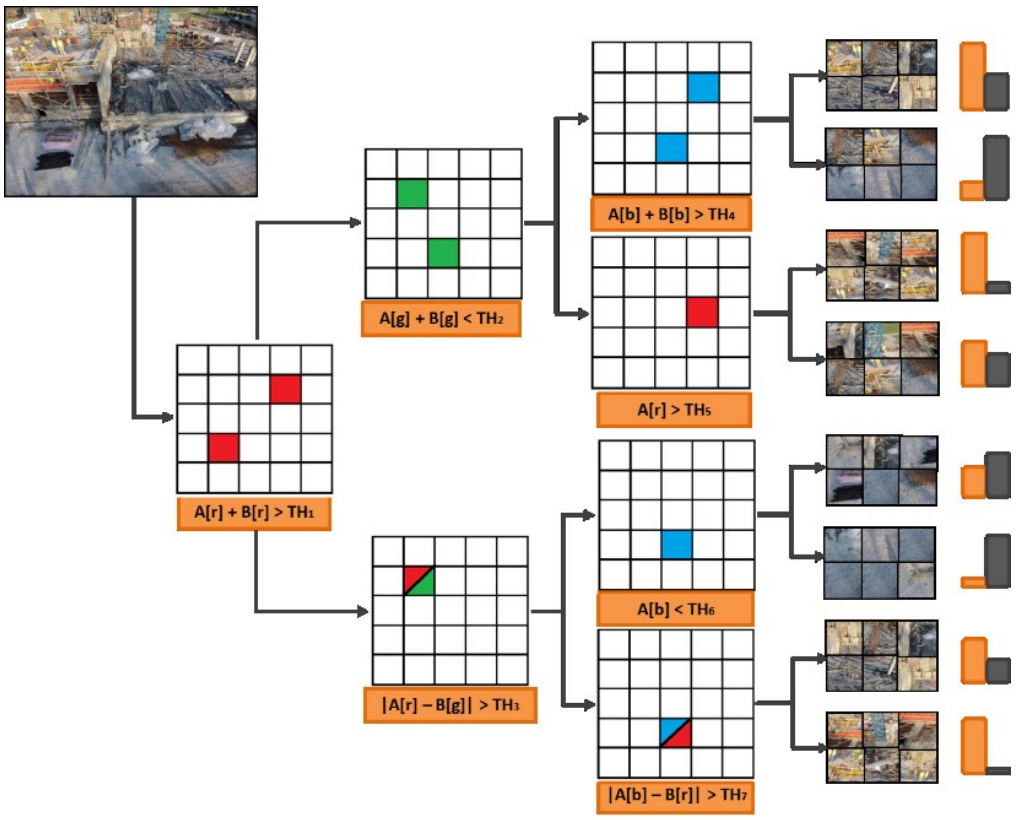


Figure 2. Schematic overview of pixel labeling based on the first-level decision tree. The threshold values of TH_1, TH_2, …, TH_n are obtained during the training process.

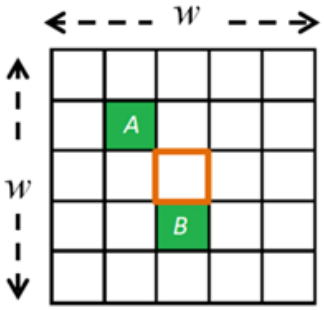
To train the first level decision tree, training images consisting of P sets of pixels are sampled. First, locations of sampling points are determined such that they are laid out in a grid pattern on the image. Grid sampling enables to reduce the computational time to generate decision trees. Patches with the size of $w \times w$ are then formed at each sampling location. For the patches, test splits are produced and randomly selected to categorize image pixels (Figure 3). Z0 and Z1 correspond to one of red, green, and blue channels, selected at random. For two randomly selected pixels A and B within the patch, $A[Z_0]$ and $B[Z_1]$ are intensity values at a corresponding color channel ranging from 0 to 255. Decision trees in a patch are constructed through dividing pixels recursively into P_{true} and P_{false}, where P_{true} is a set of pixels that satisfy the decision rule of a node and P_{false} is a set that does not satisfy it. Among randomly selected test splits for each node, a single

223 test split that maximizes the information gain at a node is selected as the decision rule of the node
224 [24,34]. An information gain at each node is calculated by the following equation:
225

$$\Delta \mathrm{E} = -\frac{|P_{true}|}{|P|} E(P_{true}) - \frac{|P_{false}|}{|P|} E(P_{false}) \quad (2)$$

226

Test Split ID	Test Split Description
1	$A[Z_0]$
2	$\mathrm{Log}(A[Z_0])$
3	$A[Z_0] + B[Z_1]$
4	$A[Z_0] - B[Z_1]$
5	$\mathrm{abs}(A[Z_0] - B[Z_1])$
6	$A[Z_0] \times \mathrm{Log}(B[Z_1])$
7	$A[Z_0] \times B[Z_1]$
8	$A[Z_0] / B[Z_1]$



227

228 Figure 3. Test splits used as low-level features (left). A patch containing sample pixels (right).
229

230 A first-level decision forest containing 5 decision trees with the depth of 10 is constructed
231 using 400 test splits at node levels; patches of 15 by 15 pixels are generated with the center-to-
232 center distance of 5 pixels. The contribution of test splits at each node are examined in the decision
233 forest, and those with high contribution are used to construct the final decision trees. In our case
234 studies, it was observed that the elimination of the test splits with minor contributions reduce the
235 training time from 253 to 177 mins by selectively focusing on highly contributed test split
236 candidates for each decision node. In this paper, the test splits with the IDs of 1, 3, 4, and 5 are
237 chosen for decision tree generation (Figure 3) as they show a great contribution on forming
238 decision trees. To prevent the overfitting problem, the decision trees are trained in a randomized
239 fashion. This randomization in the training phase is satisfied through two main criteria: (1) a subset
240 of the training pixels for each category is randomly selected; and (2) candidate decision nodes are

241 selected at random from numerous test splits to branch out the split nodes. The minimal size of
242 decision trees is selected under a predefined measure to assure that the data are thoroughly split in
243 an optimized fashion [25]. Parameters that affect the construction and performance of the decision
244 trees are: (1) a type of the split tests (Figure 3) which has a significant role in training the decision
245 trees and its performance; (2) the number of the decision trees in the forest, relevant to a trade-off
246 between precision and computational cost; (3) the color channels of images and how different
247 channels are chosen to construct the tree; (4) the maximum depth of the trees (the deeper trees can
248 differentiate more detailed visual clues but they are likely to overfit to a training dataset); and (5)
249 the size of the patch studied around each pixel. As the hyper parameters rely upon the characteristic
250 of the training dataset, we have performed trial and error tests to empirically determine the
251 parameters leading to high semantic segmentation accuracy.

252 We build the second-level decision trees based on the rectangular-based Haar features [35],
253 and the rectangular sum features [36] to obtain accurate boundaries of objects of interest. In this
254 study, we consider construction resources (e.g., equipment or materials) and built environments
255 under construction as objects of interest in jobsites. In addition, sky, soil and vegetation are
256 considered as environmental objects that are not relevant to construction monitoring.

257

258 ***4.2 Scoring the Outcome of the Semantic Segmentation***

259 Once the semantic segmentation algorithm is trained, we score the identified objects based on their
260 locations in the frames (i.e., a higher score is assigned when an object is closer to the center of a
261 frame). For this propose, we construct a 2D filter (Figure 5d) that has the same size of the original
262 image with higher pixel intensities in the center and lower intensities at the corners. The
263 algorithmic structure is shown in Figure 4. We then binarize the outcome of the semantic

segmentation by assigning 1 to pixels representing objects of interest and 0 for the rest (Figure 5c). The binarization and the convolution with the 2D filter is performed to calculate an image score. Such score is normalized with respect to the summation of all pixel intensities in the 2D filter. These operations assure (1) assigning higher scores to the image frames containing objects of interest in the middle; and (2) the independence of the scoring system with respect to the image size through the normalization.

Input: Image dimension: $n_{1\times1}$, $m_{1\times1}$
Output: Gray scale filter: Z

1 $Z \leftarrow$ create a zero array with the dimension of $(n \times m)$
2 $np_{(n\times1)} \leftarrow$ create a column vector for n evenly spaced points between -50 and 50.
3 $mp_{(1\times m)} \leftarrow$ create a row vector for m evenly spaced points between -50 and 50.
4 $npc_{(n\times m)} \leftarrow$ concatenate m number of np array to form an array with the dimension of $(n \times m)$
5 $mpc_{(n\times m)} \leftarrow$ concatenate n number of mp array to form an array with the dimension of $(n \times m)$
6 $S \leftarrow (npc \otimes npc + mpc \otimes mpc)^{0.5}$
7 for $p = 1$ **to** 5
8 **for** each element in S array located at i'th row and j'th column.
9 E ← get value of i'th row and j'th column of S array
10 **if** $E < 50 - p \times 5$
11 add value of 0.2 to the value of i'th row and j'th column of Z array
12 **end if**
13 **end for**
14 end for
15 Return Z array

The symbol ⊗ is an element-wise operation between two arrays

Figure 4: Algorithmic structure to generate the 2D filter for pixelwise multiplication.

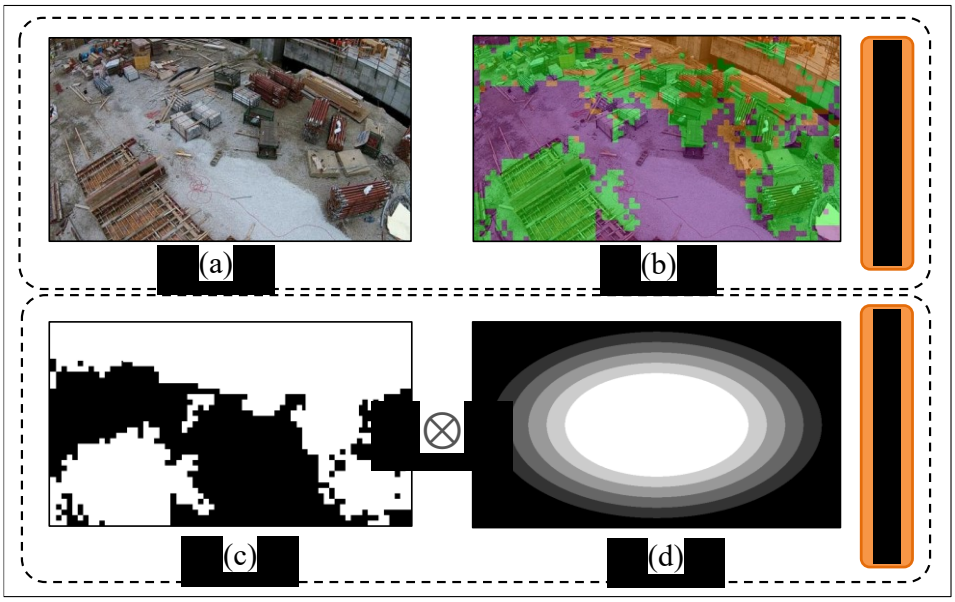


Figure 5: (a) Original image frame; (b) the sematic segmentation of the image frame; (c) Isolation of the segmented objects of interest in the binary domain; (d) 2D digital filter for pixel-wise multiplication between (c) and (d).

5. COMPOSITIONAL SIMILARITY-BASED ESTIMATION OF SNAP POINT LIKELIHOOD

For the image frames selected through the previous section, the probability of whether they were intentionally taken or not is estimated to improve the performance of visual data filtering by retrieving more meaningful image frames in the context of construction documentation. Detection of intentionally taken (or snap point) photos involves challenges to overcome. First, general snap point images do not share specific visual shapes or contents. Therefore, contents of such snap point images involve a great number of objects which demands massive dataset to carry out the training. Obtaining such training dataset is not trivial, requiring human judgment to manually select snap point images scattered throughout given visual dataset. Second, snap point images could be taken from numerous perspectives. For instance, they could be either ground-level, aerial oblique or aerial nadir images. This variety in perspective would cause poor matches between images of

291 training and testing domains in case that each is taken from different perspectives. Third, testing
292 and training domains may not necessarily involve images with the same quality. The training
293 dataset for keyframe selection techniques needs to cover large sets of objects and categories, and
294 with this reason, such datasets have been primarily collected from the internet. However, such
295 images are mostly resized or compressed to be quickly uploaded in the web, and thus they typically
296 have limited quality. The quality difference between images would reduce the matching accuracy
297 across the two domains. To address these challenges, building upon the snap point detection of
298 [4], we investigate the quality of the retrieved image frames in terms of their compositional
299 similarities comparing with the SUN dataset (~130,000 snap point images captured from various
300 perspectives), in order to select intentionally-taken photos for construction documentation. Our
301 contribution is as follows: (1) analyzing the effect of visual features in the snap point detection
302 that could be effective for filtering construction visual data from aerial perspective; (2) tuning
303 hyper-parameters to achieve optimal numbers of eigenvectors in the concatenation array by
304 performing the principal component analysis on the feature space; and (3) testing the filtering
305 performance with different threshold values to robustly filter large-scale visual data taken from
306 jobsites. The following sections elaborate the algorithmic enhancement to select photo-worthy
307 image frames collected from construction sites.
308

309 ***5.1 Feature Extraction for Snap Point Detection***

310 To compare the similarities between discriminative features, we extract Dense-SIFT [37], HOG
311 [38], GIST [39], SSIM [40], and motion blur features for each image, with the dimensions of 4000,
312 6300, 512, 6300, and 1, respectively. Since most construction images collected from aerial
313 perspective lack a visible horizon line [41] as shown in Figure 6, we do not use the horizon line
314 feature that may reduce the accuracy of the algorithm in this research, as opposed to the previous

work [4]. After extracting features, their variances in the feature space are measured through the Principal Component Analysis (PCA) to form eigenvectors for selecting the important feature dimensions that represent the input data. The PCA is a statistical operation to convert correlated variables into sets of uncorrelated variables called principal components [27,42,43]. These principal components correspond to the eigenvectors that have largest eigenvalues in dimension. The PCA reduces the size of features and thus speeds up the calculations. After obtaining the eigenvectors, a group of them showing higher variances are selected and concatenated for each type of feature to form a concatenation array. The overall accuracy of the algorithm is highly affected by the number of eigenvectors of each feature type in the concatenation array. We have empirically selected the number of eigenvectors through multiple trials. At each trial, a precision-recall curve was plotted, and the number of eigenvectors that yield the best performance were selected.

Figure 6: Absence of a horizon line in aerial visual data (bottom) obtained from construction sites.

To determine whether an image was intentionally taken or not (i.e., the intentionality of photographing), we leverage the SUN dataset [44] as the benchmark dataset, which contains ~130,000 intentionally taken images for 4,479 object categories. The role of the SUN dataset is to

provide the filtering criteria. After extracting features and applying the PCA for visual data from construction sites and the SUN dataset, the compositional similarities across the two domains are computed. The similarity computation is discussed in the following sections with more details.

5.2 Domain Adaptation and Computing Similarity

We seek a domain invariant feature space to increase the probability of a valid match between the benchmark and the testing datasets. We build an infinite dimensional geodesic path to connect the two domains through common feature subspaces. This path would substantially decrease the number of mismatches caused by camera resolutions and differences between the two domains [42]. The geodesic path can be expressed as follows:

$$K_{GFK}(x_i, x_j) = (z_i^{\infty}, z_j^{\infty}) = \int_0^1 (\Phi(t)^T x_i)^T (\Phi(t)^T x_j) dt \quad (3)$$

The subspaces along the geodesic path is denoted by $\Phi(t)$. x_i, x_j, which represent the visual features of the benchmark and testing datasets respectively. $z_i^{\infty}, z_j^{\infty}$ denote the infinite dimension containing all of projections of x_i, x_j along the geodesic path. The distance of projections which transfers from the benchmark domain to the testing domain is denoted by t. Once the geodesic path is calculated, indefinite sets of projections are generated via Geodesic Flow Kernel (GFK).

5.3 Snap Point Prediction

A set of images in the benchmark dataset could be retrieved for any testing images that have a higher similarity, and particularly those that have higher GFK values. Let's assume that we retrieve *k* number of photos in the training domain with the highest GFK values.

$$S(x^e) = \sum_{j=1}^{k} K_{GFK}\left(x^e,\ x^w{}_j\right) \quad (4)$$

356

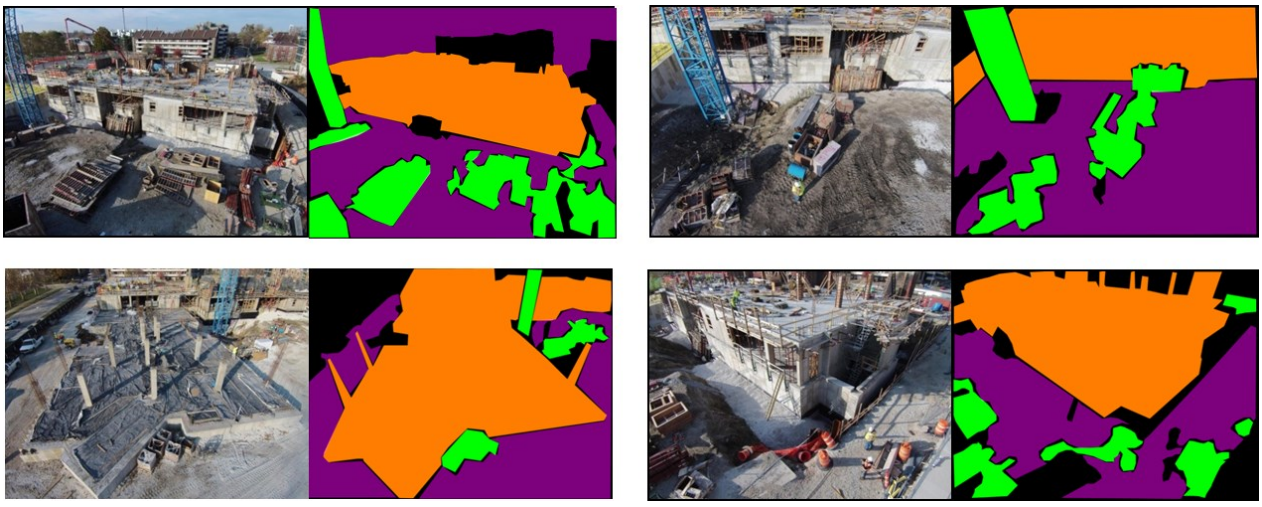
357 where, $x^e and\ x^w$ denotes image descriptors for a testing image and a benchmark image,
358 respectively. k is the number of the retrieved similar images from the benchmark data set, and
359 $S(x^e)$ serves as the confidence of snap points of the testing image frame. The higher values of
360 $S(x^e)$ indicates the higher chance of a testing frame taken intentionally. With all these parameters,
361 we predict snap point frames for any given footages. For each image, $S(x^e)$ is derived and images
362 are sorted out from the lowest score to the highest one. In this paper, we categorize images in three
363 groups depending on the extent to which the photos were taken intentionally. To achieve this
364 categorization, as a proof of concept, we leveraged two reference images to serve as discriminative
365 boundaries between three groups of images. These two images (each with a different level of
366 photo-worthy strength) were taken from a construction site using a UAV. They are manually
367 selected and used for two purposes: 1) to act as discriminative boundaries between different groups
368 of images; and 2) to serve as the scoring reference while obtaining the ground truth scores for the
369 testing images. We derive scores for these two reference images as we did for all testing images.
370 Then, testing images with lower scores than the first reference image are considered as the first
371 group, images within the range of scores of the two reference images are given to the second group,
372 and the remaining images with higher scores than the score of the second reference image are
373 assigned to the last group. Scores for all of testing images are assigned depending on their snap
374 point confidence index $S(x^e)$.

375

6. CASE STUDIES AND EVALUATIONS

6.1. Experimental Setup (ground-truth data)

The training images for the semantic segmentation were collected using a UAV from different stages of multiple construction sites; those images taken from various viewpoints contain built environments under construction, construction resources, sky and ground soil, and vegetation. The collected images had dimensions that varies from 852 × 450 to 4384 × 3288. They were resized to 600 pixels in width, and the height was proportionally adjusted, while preserving the original aspect ratio. We used 90 images for training, and 29 images for testing the performance of the semantic segmentation. To obtain labeled pixel level classes, we have manually annotated the boundaries of different categories within the training and testing images (Figure 7). The outcome of pixel-level labeling is an image in which each class is separated by each region with a specific color. The semantic categories of target objects were built environment under construction (orange), construction resources (green), and environmental objects (purple). The black (NULL) category shown in the training images represents unlabeled regions, and therefore those regions were not used during the evaluation of the proposed method.



393 Figure 7: Examples of annotated categories in images used as ground truth for training and testing.

394

395 To obtain the ground truth data for assessing the performance of the proposed method,
396 participants were invited to assign scores to the given testing images. Scoring the photo-worthiness
397 to images is an inevitably subjective task since there is no absolute standard in terms of the
398 worthiness. In this respect, the photo-worthiness of construction site images is a relatively
399 conceptual and empirical value from individual to individual by nature. Here, people are likely to
400 share the similar point of view when they evaluate the quality of images with respect to a particular
401 application. For example, construction practitioners can judge which image frames are valuable
402 and which are not based on their reasoning process, and some image frames can be acknowledged
403 as valuable by most construction practitioners for a certain purpose of project management. Based
404 on this fact, in our case studies, the ground truth data were obtained by different participants who
405 have professional experiences in the construction domain, and multiple scoring results were
406 averaged to obtain consistent scores over the entire testing data. To obtain reliable scores for the
407 images, two strategies were implemented: 1) similar to the previous work of snap point detection
408 [4], the score of each image was averaged by scores given by different participants; and 2) two
409 reference images (each with a score) were provided to the participants to assist their scoring tasks.
410 Specifically, we provided two reference images taken from jobsites to the participants; the
411 characteristics of the two images were as follows: 1) the first image with the score of 50 partially
412 contains the objects of interest and was relatively intentionally taken; and 2) the second image with
413 the score of 80 was intentionally taken and the middle of the image was mainly occupied by
414 construction-related contents.

415

6.2. Tuning Hyperparameters and Training

For constructing the first level decision trees, the hyper parameters of the semantic random forest, such as the number of decision trees, the number of test splits at each node, and the depth of decision trees, were studied to obtain the high accuracy and computational efficiency (Figure 8). The computational efficiency for the semantic segmentation was measured with a consumer-level engineering workstation composed of an i7 CPU and an NVIDA GeForce GT 550M graphic card. The global accuracy for each image was measured in the form of the percentage of pixels correctly classified in the semantic segmentation.

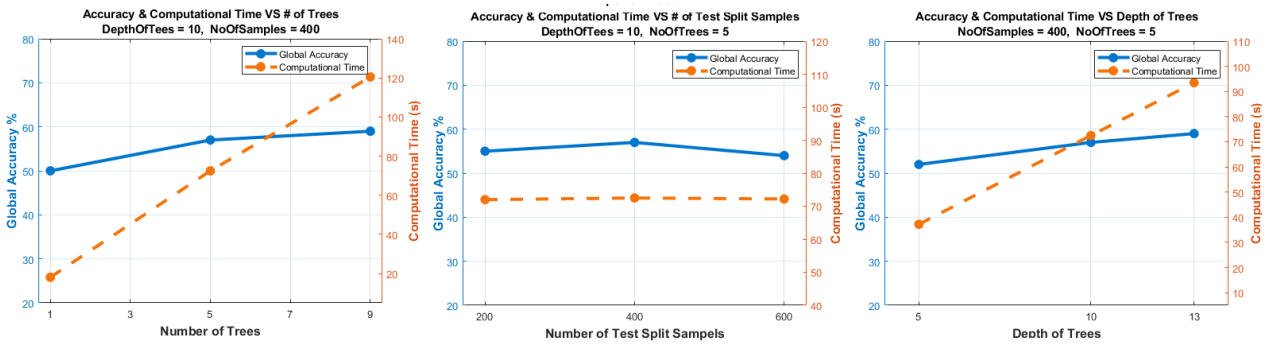


Figure 8: The effect of semantic segmentation parameters on accuracy and computational cost

As can be seen in Figure 8, it was observed that increasing the number and the depth of decision trees entails the additional computational cost but did not lead to a significant accuracy improvement. Furthermore, in our case studies, we observed 400 test splits have yielded the highest accuracy with almost the same computational cost comparing to 200 and 600 test splits. We have deployed five sets of decision trees to perform the decision splits for a square patch with the size of 15 by 15 pixels. The distance between patches was set to 5 pixels. To calculate test splits at each node, 400 scenarios were set at each patch and the depth of trees were set to 10 in

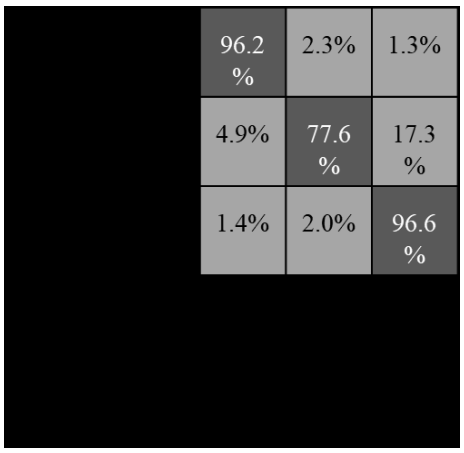
435 our case studies.

436

437 **6.3. Experimental Results and Discussion**

438 In order to evaluate the performance of the proposed method in the real-world and to facilitate the
439 transfer of high-quality evidence from our research into practice, we have conducted the case
440 studies. We have collected visual data from two real-world construction sites where construction
441 projects were being progressed (Case #1: a student housing building construction project in
442 Illinois, with 800 regularly sampled images; and Case #2: an instructional facility construction
443 project in Maryland, with 304 regularly sampled images). The reason why we selected the specific
444 cases above is that the two cases represent typical mid-scale construction projects and provide
445 general scenes of building construction as shown in Figure 7. Aerial visual data from such
446 representative cases will provide an excellent opportunity to test the proposed method in terms of
447 broad applicability. The collected visual dataset using drones was composed of photo-worthy
448 images reflecting the as-is status of the jobsites and accidentally taken images which most scenes
449 are occupied by ground or sky other than the jobsites. To reduce the computational cost for visual
450 data filtering, we resized the images to be 600 pixels wide, and regularly sampled a frame in every
451 0.2 to 1.5 seconds in the entire videos based on the similarity level between adjacent video frames.
452 The confusion matrix was generated to report the segmentation accuracy (Figure 9). In the
453 experiments, the average accuracy for three classification categories was 90.1%. Figure 10 shows
454 examples of the semantic segmentation regarding the two case studies (Orange: built environment
455 under construction, Green: construction resource, Purple: environmental objects consisting of sky,
456 ground soil, and vegetation). We observed that there was a relatively high confusion between
457 'Construction Resources' and 'Environmental Objects' categories as shown in Figure11.

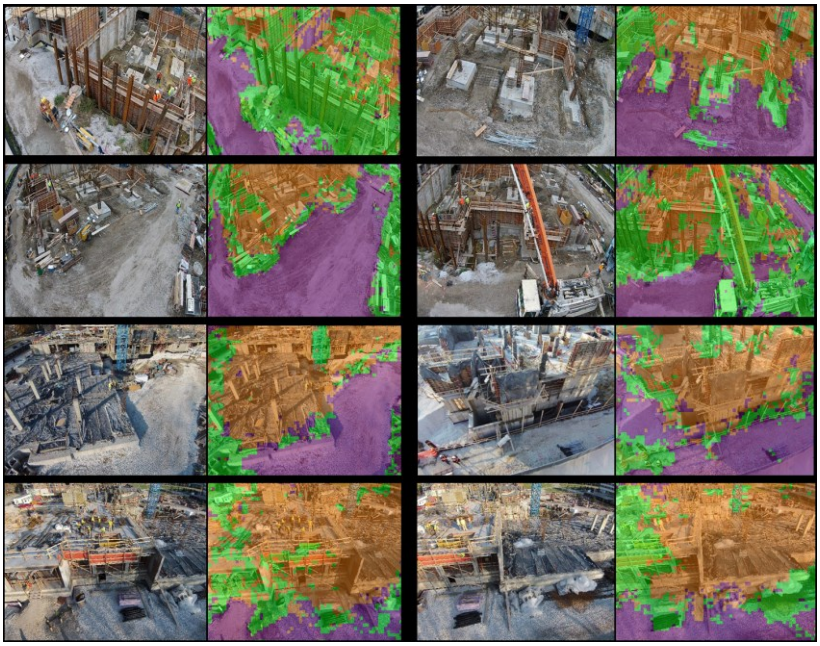
458



459

460 Figure 9: Confusion matrix of the semantic segmentation in our case study.

461



462

463 Figure 10: Examples of sematic segmentation for different construction jobsites.

464

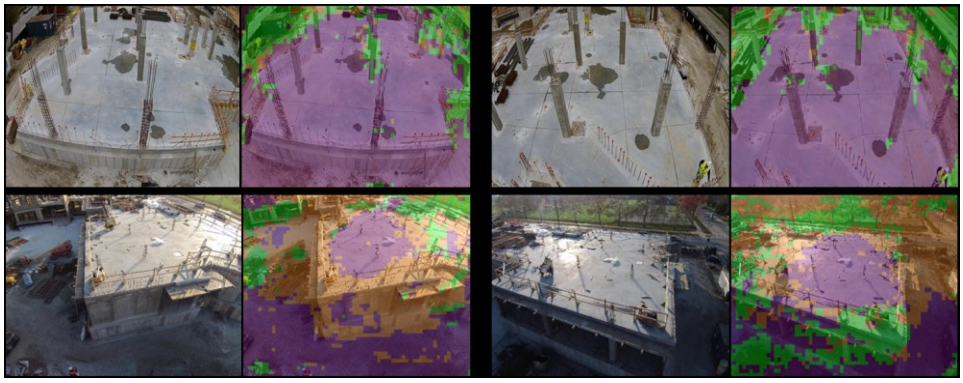


Figure 11: Examples of poor segmentation between semantic categories.

Through the semantic segmentation stage, image frames involving objects of interest that appear in the middle of the image frame were given high scores. The scores were normalized, and the images with the score of 80 or more were retrieved in our case studies. Then, in the second filtering step, snap point values (i.e., photo-worthiness) of testing images were obtained and normalized, and those with score of 80 or more were considered as photo-worthy images. The performance of the proposed method was compared with the snap point detection method [4] that retrieves intentionally-taken photos, and the saliency method [3,45] that investigates the presence of objects aesthetically showing good quality. For the comparison, a precision-recall curve was obtained to evaluate the overall performance of visual data filtering with different threshold values at each filtering process. To obtain precision and recall values, True Positive (TP), False Positive (FP) and False Negative (FN) values were calculated at each threshold. TP is the number of images that are correctly classified as photo-worthy, FP is the number of images that are incorrectly classified as photo-worthy, and FN is the number of photo-worthy images that are not classified as photo-worthy. Precision and recall values are calculated as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

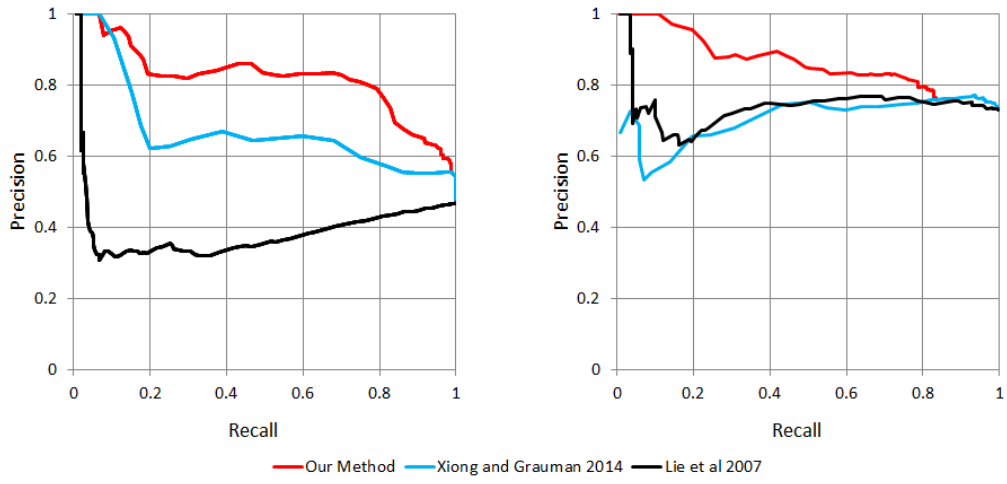
483

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

484

485 In our case studies, a threshold value at the second filtering process was fixed as 80 in the
486 comparison study, and then different precision and recall values were obtained by changing a
487 threshold value at the semantic segmentation stage from 0 to 100. Likewise, the precision-recall
488 curves of the prior works of the snap point detection [4] and the saliency method [3,45] were
489 obtained by changing threshold values of their scores to select photo-worthy images. Figure 12
490 shows the obtained precision-recall curves.

491



492

493 Figure 12: Precision-recall curves of our method (red), (Xiong and Grauman 2014) (blue), and
494 (Dhar et al 2011) (black) for filtering raw video footages taken from aerial perspective in
495 construction jobsites (Left: Case#1, Right: Case #2).

496

497 In our experiments, we observed that the proposed method delivers a close-to-human
498 judgment for photo selection from raw video footages in the context of construction

documentation, as shown in Figure 13. There are two possible reasons why it was observed that the proposed method performed better than the prior works, in the context of construction documentation. First, the proposed method employed the semantic segmentation at the first filtering process to filter out images without construction-related objects. Moreover, the composition of visual contents was also considered to evaluate the photo-worthiness based on the locations of construction related-objects in an image frame. We believe that this strategy militated in favor of the selection of construction-related photo-worthy images. Second, we have investigated and used the highly influential visual features in the context of construction documentation at the second filtering stage, which is different from others in the context of daily life [4]. Such feature engineering was found to be helpful for improving the filtering performance in the context of construction site monitoring.

Figure 13: Examples of the experimental results of the proposed method. Top: unlikely snap points, and Bottom: 'photo-worthy' frames.

Deploying computational methods to perform visual data filtering for achieving a higher level of accuracy (especially, the part of semantic segmentation) generally require additional

computational cost as well as a large number of semantically labeled images as a training dataset. Rather, we have mainly focused on algorithmic improvement for filtering construction-related large-scale visual data building upon reasonably fast and yet robust methods that can be performed with the typical computation power of consumer-level computers possessed by practitioners. In this regard, we built upon low-level features to perform faster both in the training and testing phases and tuned hyper parameters in a way that the computational cost can be reduced while the algorithm performs with a reasonable range of accuracy. Overall, through our case studies, we could observe that the proposed method promises the potential of selecting photo-worthy image frames related to construction projects from raw visual data.

7. CONCLUSION

The proliferation of affordable camera-equipped devices that can record high quality video footages promotes visual monitoring of construction sites. As a result, large-scale visual data in the form of numerous still images or long-sequence video sequences are being collected from construction sites on daily, weekly or monthly bases. In this regard, the visual data overload is regarded as a major issue in visual monitoring of construction projects. To address such challenge, this paper presents a new method to filter construction visual data building upon (1) semantically segmenting contents of images to investigate the presence of objects of interest and its spatial composition; and (2) estimating and assessing the likelihood of the image frames to be intentionally taken. In our two case studies, the proposed method has shown the high accuracy in identifying photo-worthy image frames from large-scale raw visual dataset for construction monitoring purposes that is close to visual data analytics based on human judgment. The main contribution of this work is to automate construction visual data filtering, retrieving images that

are valuable for documenting the as-is status of construction jobsites from video footages captured via UAVs. Such automation will save time and effort of practitioners who should manually select and document informative image frames to report or assess various construction performance metrics (e.g., construction progress or safety). While this work has demonstrated the potential to convert large-scale visual dataset into concise visual photologs in the construction domain, open research challenges still exist: (1) as a proof of concept, limited numbers of object categories were studied in the semantic segmentation process of our case studies. Segmentation with more categories will retrieve more distinct objects from given visual dataset, but require larger training dataset that would be computationally expensive. Further study needs to investigate such trade-off for robust visual data filtering in the construction domain; and (2) to minimize the overlap among filtered image frames, raw video frames were sampled based on the sampling frequency that was empirically determined in our case studies. Here, the high sampling frequency may need to be considered for particular locations in construction sites that requires rigorous monitoring and documentation. In this case, it would be better to set an adaptive sampling frequency for visual dataset representing those areas, and thus more research needs to be conducted to study such dynamic sampling rate with respect to image contents for effective construction documentation. All these issues are currently being explored as part of our ongoing research.

ACKNOWLEDGEMENT

This material is in part based upon work supported by the National Science Foundation (NSF) under CMMI Award #1832187. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

563 **REFERENCES**

564 [1] K.K. Han, M. Golparvar-Fard, Potential of big visual data and building information
565 modeling for construction performance analytics: An exploratory study, Automation in
566 Construction. 73 (2017) 184–198. https://dx.doi.org/10.1016/j.autcon.2016.11.004.
567 [2] Y. Furukawa, J. Ponce, Accurate, Dense, and Robust Multiview Stereopsis, IEEE
568 Transactions on Pattern Analysis and Machine Intelligence. 32 (2010) 1362–1376.
569 https://dx.doi.org/10.1109/TPAMI.2009.161.
570 [3] S. Dhar, V. Ordonez, T.B.-C.V. and Pattern, U. 2011, High level describable attributes for
571 predicting aesthetics and interestingness, in: Computer Vision and Pattern Recognition
572 (CVPR), 2011 IEEE Conference On, . https://dx.doi.org/10.1109/CVPR.2011.5995467.
573 [4] B. Xiong, K. Grauman, Detecting Snap Points in Egocentric Video with a Web Photo Prior,
574 in: Computer Vision -- ECCV 2014: 13th European Conference, Zurich, Switzerland,
575 September 6-12, 2014, Proceedings, Part V, Springer International Publishing, Cham, 2014:
576 pp. 282–298. https://dx.doi.org/10.1007/978-3-319-10602-1_19.
577 [5] M. Chen, S. Mao, Y. Liu, Big Data: A Survey, Mobile Networks and Applications. 19
578 (2014) 171–209. https://dx.doi.org/10.1007/s11036-013-0489-0.
579 [6] A. Hanjalic, Extracting moods from pictures and sounds: towards truly personalized TV,
580 IEEE Signal Processing Magazine. 23 (2006) 90–100.
581 https://dx.doi.org/10.1109/MSP.2006.1621452.
582 [7] R.M. Jiang, A.H. Sadka, D. Crookes, Advances in Video Summarization and Skimming,
583 in: Springer, Berlin, Heidelberg, 2009: pp. 27–50. https://dx.doi.org/10.1007/978-3-642-
584 02900-4_2.
585 [8] A. Rashidi, F. Dai, I. Brilakis, P. Vela, Optimized selection of key frames for monocular
586 videogrammetric surveying of civil infrastructure, Advanced Engineering Informatics. 27
587 (2013) 270–282. https://dx.doi.org/10.1016/j.aei.2013.01.002.
588 [9] M. Gygli, H. Grabner, L. Van Gool, Video summarization by learning submodular mixtures
589 of objectives, in: Proceedings of the IEEE Conference on Computer Vision and Pattern
590 Recognition, 2015: pp. 3090–3098. https://dx.doi.org/10.1109/CVPR.2015.7298928
591 [10] Y.S. Avrithis, A.D. Doulamis, N.D. Doulamis, S.D. Kollias, A Stochastic Framework for
592 Optimal Key Frame Extraction from MPEG Video Databases, Computer Vision and Image
593 Understanding. 75 (1999) 3–24. https://dx.doi.org/10.1006/CVIU.1999.0761.

[11] A. Khosla, R. Hamid, C.-J. Lin, N. Sundaresan, Large-Scale Video Summarization Using Web-Image Priors, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2013: pp. 2698–2705. https://dx.doi.org/10.1109/CVPR.2013.348.

[12] V. Chandrasekhar, W. Min, X. Li, C. Tan, B. Mandal, L. Li, J. Hwee Lim, Efficient Retrieval from Large-Scale Egocentric Visual Data Using a Sparse Graph Representation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014: pp. 527–534. https://dx.doi.org/10.1109/CVPRW.2014.84.

[13] G. Kim, L. Sigal, E.P. Xing, Joint Summarization of Large-Scale Collections of Web Images and Videos for Storyline Reconstruction, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014: pp. 4225–4232. https://dx.doi.org/10.1109/CVPR.2014.538.

[14] Z. Lu, K. Grauman, Story-Driven Summarization for Egocentric Video, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013: pp. 2714–2721. https://dx.doi.org/10.1109/CVPR.2013.350.

[15] M. Gygli, H. Grabner, H. Riemenschneider, L. Van Gool, Creating Summaries from User Videos, In European conference on computer vision: Springer, Cham, 2014: pp. 505–520. https://dx.doi.org/10.1007/978-3-319-10584-0_33.

[16] J. Kopf, M.F. Cohen, R. Szeliski, First-person hyper-lapse videos, ACM Transactions on Graphics. 33 (2014) 1–10. https://dx.doi.org/10.1145/2601097.2601195.

[17] T. Liu, J.R. Kender, Optimization Algorithms for the Selection of Key Frame Sequences of Variable Length, in: European Conference on Computer Vision, 2002: pp. 403–417. https://dx.doi.org/10.1007/3-540-47979-1_27.

[18] D. Liu, Gang Hua, Tsuhan Chen, A Hierarchical Visual Model for Video Object Summarization, IEEE Transactions on Pattern Analysis and Machine Intelligence. 32 (2010) 2178–2190. https://dx.doi.org/10.1109/TPAMI.2010.31.

[19] M. Spain, P. Perona, Measuring and Predicting Object Importance, International Journal of Computer Vision. 91 (2011) 59–76. https://dx.doi.org/10.1007/s11263-010-0376-0.

[20] Y.J. Lee, J. Ghosh, K. Grauman, Discovering important people and objects for egocentric video summarization, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference On, 2012: pp. 1346–1353. https://dx.doi.org/10.1109/CVPR.2012.6247820.

[21] T.L. Berg, D.A. Forsyth, Animals on the Web, in: Computer Vision and Pattern

625 Recognition, IEEE Computer Society Conference On, IEEE, 2006: pp. 1463–1470.
626 https://dx.doi.org/10.1109/CVPR.2006.57.
627 [22] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, J.M. Rehg, A Scalable Approach to
628 Activity Recognition based on Object Use, in: 2007 IEEE 11th International Conference on
629 Computer Vision, 2007: pp. 1–8. https://dx.doi.org/10.1109/ICCV.2007.4408865.
630 [23] A. Dimitrov, M. Golparvar-Fard, Vision-based material recognition for automated
631 monitoring of construction progress and generating building information modeling from
632 unordered site image collections, Advanced Engineering Informatics. 28 (2014) 37–49.
633 https://dx.doi.org/10.1016/j.aei.2013.11.002.
634 [24] M. Johnson, J. Shotton, Semantic Texton Forests, in: Computer Vision: Detection,
635 Recognition and Reconstruction, 2010: pp. 173–203. https://dx.doi.org/10.1007/978-3-642-
636 12848-6_7.
637 [25] J. Shotton, M. Johnson, R. Cipolla, Semantic texton forests for image categorization and
638 segmentation, in: Computer Vision and Pattern Recognition, 2008 IEEE Conference On,
639 IEEE, 2008: pp. 1–8. https://dx.doi.org/10.1109/CVPR.2008.4587503.
640 [26] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder
641 architecture for image segmentation, IEEE Transactions on Pattern Analysis and Machine
642 Intelligence. 39 (2017) 2481–2495. https://dx.doi.org/10.1109/TPAMI.2016.2644615.
643 [27] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometrics and
644 Intelligent Laboratory Systems. 2 (1987) 37–52. https://dx.doi.org/10.1016/0169-
645 7439(87)80084-9.
646 [28] J. Gong, C.H. Caldas, Learning and Classifying Motions of Construction Workers and
647 Equipment Using Bag of Video Feature Words and Bayesian Learning Methods, in:
648 Computing in Civil Engineering (2011), American Society of Civil Engineers, Reston, VA,
649 2011: pp. 274–281. https://dx.doi.org/10.1061/41182(416)34.
650 [29] L. Fei-Fei, P. Perona, A bayesian hierarchical model for learning natural scene categories,
651 in: Computer Vision and Pattern Recognition, 2005 IEEE Computer Society Conference
652 On, 2005: pp. 524–531. https://dx.doi.org/10.1109/CVPR.2005.16.
653 [30] R. Szeliski, Computer vision: algorithms and applications, 2010.
654 [31] M.-W. Park, I. Brilakis, Construction worker detection in video frames for initializing vision
655 trackers, Automation in Construction. 28 (2012) 15–25.

https://dx.doi.org/10.1016/j.autcon.2012.06.001.

[32] M. Memarzadeh, M. Golparvar-Fard, J.C. Niebles, Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors, Automation in Construction. 32 (2013) 24–37. https://dx.doi.org/10.1016/J.AUTCON.2012.12.002.

[33] A. Heydarian, M. Memarzadeh, M. Golparvar-Fard, Automated Benchmarking and Bonitoring of an Earthmoving Operation's carbon footprint using video cameras and a greenhouse gas estimation model, in: Computing in Civil Engineering (2012), 2012: pp. 509–516. https://dx.doi.org/10.1061/9780784412343.0064.

[34] V. Lepetit, P. Lagger, P. Fua, Randomized Trees for Real-Time Keypoint Recognition, in: Computer Vision and Pattern Recognition, 2005 IEEE Computer Society Conference On, IEEE, : pp. 775–781. https://dx.doi.org/10.1109/CVPR.2005.288.

[35] P. Viola, M. Jones, Rapid Object Detection Using a Boosted Cascade of Simple Features, in: Computer Vision and Pattern Recognition, 2001 IEEE Conference On, : p. I-511-I-518. https://dx.doi.org/10.1109/CVPR.2001.990517.

[36] J. Shotton, J. Winn, C. Rother, A. Criminisi, TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context, International Journal of Computer Vision. 81 (2009) 2–23. https://dx.doi.org/10.1007/s11263-007-0109-1.

[37] A. Vedaldi, B. Fulkerson, Vlfeat, in: Proceedings of the International Conference on Multimedia - MM '10, ACM Press, New York, New York, USA, 2010: p. 1469. https://dx.doi.org/10.1145/1873951.1874249.

[38] N. Dalal, B. Triggs, Histograms of Oriented Gradients for Human Detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), IEEE, : pp. 886–893. https://dx.doi.org/10.1109/CVPR.2005.177.

[39] A. Oliva, A. Torralba, Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, International Journal of Computer Vision. 42 (2001) 145–175. https://dx.doi.org/10.1023/A:1011139631724.

[40] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multiscale structural similarity for image quality assessment, in: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, IEEE, : pp. 1398–1402.

687 https://dx.doi.org/10.1109/ACSSC.2003.1292216.
688 [41] J. Košecká, W. Zhang, Video Compass, in: European Conference on Computer Vision,
689 2002: pp. 476–490. https://dx.doi.org/10.1007/3-540-47979-1_32.
690 [42] B. Gong, Y. Shi, F. Sha, K. Grauman, Geodesic Flow Kernel For Unsupervised Domain
691 Adaptation, in: Computer Vision and Pattern Recognition, 2012 IEEE Conference On,
692 2012: pp. 2066–2073. https://dx.doi.org/10.1109/CVPR.2012.6247911.
693 [43] J. Shlens, A Tutorial on Principal Component Analysis, (2014).
694 http://arxiv.org/abs/1404.1100.
695 [44] J. Xiao, J. Hays, K.A. Ehinger, A. Oliva, A. Torralba, SUN database: Large-scale scene
696 recognition from abbey to zoo, in: Proceedings of the IEEE Computer Society Conference
697 on Computer Vision and Pattern Recognition, 2010: pp. 3485–3492.
698 https://dx.doi.org/10.1109/CVPR.2010.5539970.
699 [45] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, Heung-
700 Yeung Shum, Learning to Detect a Salient Object, IEEE Transactions on Pattern Analysis
701 and Machine Intelligence. 33 (2011) 353–367. https://dx.doi.org/10.1109/TPAMI.2010.70.
702