Multitemporal Very High Resolution From Space: 
Outcome of the 2016 IEEE GRSS Data Fusion Contest
L. Mou, X. Zhu, M. Vakalopoulou, K. Karantzalos, N. Paragios, B. Le Saux, 
G. Moser, D. Tuia

To cite this version:

HAL Id: hal-01561333
https://hal.archives-ouvertes.fr/hal-01561333
Submitted on 12 Jul 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Multitemporal Very High Resolution From Space: Outcome of the 2016 IEEE GRSS Data Fusion Contest

L. Mou, Student Member, IEEE, X. Zhu, Senior Member, IEEE, M. Vakalopoulou, Student Member, IEEE, K. Karantzalos, Senior Member, IEEE, N. Paragios, Fellow, IEEE, B. Le Saux, G. Moser, Senior Member, IEEE, and D. Tuia, Senior Member, IEEE

Abstract—In this paper, the scientific outcomes of the 2016 Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society are discussed. The 2016 Contest was an open topic competition based on a multitemporal and multimodal dataset, which included a temporal pair of very high resolution panchromatic and multispectral Deimos-2 images and a video captured by the Iris camera on-board the International Space Station. The problems addressed and the techniques proposed by the participants to the Contest spanned across a rather broad range of topics, and mixed ideas and methodologies from the remote sensing, video processing, and computer vision. In particular, the winning team developed a deep learning method to jointly address spatial scene labeling and temporal activity modeling using the available image and video data. The second place team proposed a random field model to simultaneously perform coregistration of multitemporal data, semantic segmentation, and change detection. The methodological key ideas of both these approaches and the main results of the corresponding experimental validation are discussed in this paper.

Index Terms—Change detection, convolutional neural networks (CNN), deep learning, image analysis and data fusion, multiresolution, multisource, multimodal, random fields, tracking, video from space.

I. INTRODUCTION

PUBLIC awareness about Earth observation has raised dramatically during the latest years. Among the main driving forces, we can mention the increased availability of very high spatial resolution (VHR) data acquired by sensors on-board drones, aircrafts, and satellites. These data make it possible to address new monitoring tasks, such as object detection and classification in urban areas (which are nowadays approached at a global scale using satellite acquisitions [1]), urban area classification [2], and detailed mapping at decimeter scale for precision agriculture [3]. On one hand, very high spatial resolution implies new or more complex problems, typically dealing with spatial smoothness [4], [5], misregistrations [6], changes in illumination, acquisition conditions, or sensor properties [7]–[10], and increased complexity of the signatures of the classes [11], [12]. These challenges generally raise the need for more and more sophisticated methodologies. On the other hand, the advances in hardware and the increased availability of data makes it possible to use training models that are increasingly able to extract complex and abstract features. Prominent examples are deep learning methods, which train feature extraction filters directly and nonparametrically from the data [13], [14], or feature learners, which mine the parameter spaces of existing filter banks [15], [16] or of their combinations [17].

While VHR image processing is an active field of investigation, the processing of video data taken from satellite platforms is still in its infancy. On one hand, video from space (also known as “space video”) is not currently widespread. A more common approximation is represented by sensors that can retarget a specific region and provide a multiangular sequence. For example, the hyperspectral CHRIS/PROBA sensor can provide angular sequences, which were used for chlorophyll [18] and leaf area index retrieval [18], [19], and the multitemporal multispectral WorldView-2 sensor was used for height retrieval [20], detection [21], land use classification [22], and tracking [23]. On the other hand, with the increasing development of small and cheap
commercial satellites, such as the SkySat-constellation of Skybox Imaging, wide area spaceborne remote sensing videos are becoming increasingly accessible at a reasonable cost. Indeed, while the processing of aerial videos collected by sensors on-board terrestrial vehicles or drones is not yet a major trend in remote sensing, first applications can be found in the recent literature. These applications include tracking or navigation using videos from web cameras [24], terrestrial vehicles [25], [26], aircrafts [27], and satellites [28]. Information extraction from space videos remains a major challenge, although the acquisition technology is available, as demonstrated by the Iris camera operated by Urthecast on the International Space Station (ISS).\(^1\) The Data Fusion Contest (DFC) 2016 addressed this challenge.

Since 2006, the Image Analysis and Data Fusion Technical Committee (IADF TC\(^2\)) of the IEEE Geoscience and Remote Sensing Society (IEEE-GRSS) has organized an annual DFC, in which a dataset has been released free of charge to the international community along with a data fusion competition [29]–[37]. To tackle the aforementioned challenges implied by new VHR data with very high temporal resolution, the DFC 2016 proposed a competition related to the multimodal processing of image sequences and video streams acquired from space.

The DFC 2016 released to the international community an image dataset involving VHR multiresolution and multisensor imagery and video information. The dataset was composed of a series of two VHR multispectral and panchromatic images and of a video stream acquired by the Iris camera over the city center of Vancouver, Canada (see Section II). The data were provided by the Deimos Imaging and Urthecast companies.

The competition was framed as follows: each participating team was invited to submit an original open-topic manuscript addressing an information extraction problem based on the dataset provided. All submissions were evaluated by an Award Committee, composed of the organizers of the Contest, of industrial representatives of Deimos Imaging and Urthecast, and of several past Chairs of the IADF TC. The manuscripts were ranked on the basis of scientific novelty and originality, methodological approach, experimental discussion, and quality of presentation.

In this paper, after describing the dataset of the DFC 2016 (see Section II), we will discuss first the overall scientific outcome of the contest as a whole (see Section III). Then, we will focus in more detail on the approaches proposed by the first and second place teams (see Sections IV and V, respectively). Finally, conclusions will be drawn in Section VI.

II. Dataset of the DFC 2016

The dataset of the DFC 2016 included VHR imagery and video from space covering an urban and harbor area in Vancouver, Canada (49°15'N, 123°6'W; see Fig. 1).

1) VHR Images: A temporal pair of VHR acquisitions (Deimos-2 standard products), each composed of a panchromatic image at 1-m point spacing and a multispectral image at 4-m point spacing, was included in the dataset. The two acquisition dates were March 31 and May 30, 2015. Deimos-2 operates from a Sun-synchronous orbit at a mean altitude of 620 km. The spacecraft design is based on an agile platform for fast and precise off-nadir imaging (up to ±30° over nominal scenarios and up to ±45° in emergency cases), and carries a push-broom VHR camera with five spectral channels: one panchromatic channel and four multispectral channels corresponding to red (R), green (G), blue (B), and near-infra-red (NIR) bands.

For each date, four image products were provided for the contest: the panchromatic and the multispectral images were made available at both levels 1B and 1C. A level 1B product is calibrated and radiometrically corrected but is not resampled, and geometric information is provided separately (see Fig. 2). A level 1C product is calibrated, radiometrically corrected, and manually orthorectified and resampled to a map grid, while the geometric information is contained in the GeoTIFF tags (see Fig. 3). As a consequence, the two level 1C images cover exactly the same ground area.

2) Video From Space: A full-color ultrahigh definition (UHD) video acquired by the Iris camera on-board the ISS was provided. The acquisition took place on July 2, 2015, over

---

1[https://www.urthecast.com/data#iris](https://www.urthecast.com/data#iris)

MOU et al.: MULTITEMPORAL VERY HIGH RESOLUTION FROM SPACE: OUTCOME OF THE 2016 IEEE GRSS DATA FUSION CONTEST

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Multitemporal analysis, change detection, object detection, and image classification.

III. SUBMISSIONS AND RESULTS

A. Submissions

Twenty-four papers were submitted to the DFC 2016. They proved to be of great diversity, because they addressed heterogeneous topics and proposed various methodological solutions for each topical area. Fig. 5 summarizes both these aspects.

1) Topic-wise, a majority of papers dealt with two thematic areas: Classification/change detection and tracking. While the former had been mainstream for years for the DFC, the latter emerged as a new topical area. Indeed, it also attracted most of the participants’ efforts, essentially because the data were well suited for this task. Furthermore, in addition to the two aforementioned main topics, participants also considered some other classical fusion problems (pan-sharpening and registration) as well as new topical areas such as population modeling or estimation of human activities.

2) Methodologically speaking, we observed a wide spectrum of proposals. Approaches that were quite established in remote sensing (e.g., feature extraction or sparse methods) were challenged by techniques issued from the video processing literature (e.g., optical flow and tracking) and from the deep learning and computer vision literature (e.g., convolutional neural networks (CNNs) and structured output modeling). This trend was expected because deep learning is becoming more and more prominent in the remote sensing area as well (e.g., [13] and [14]), and allows robust models to be learned even through simple fine tuning of existing architectures drawn from other methodological areas [36], [38].

Another interesting observation can be made when considering the data types used by the participating teams. Three separate sources of data (multispectral VHR, panchromatic VHR, and video) were provided but no fixed topic was imposed. Therefore, each team was free to use the most relevant sources to address their problem. Fig. 6 summarizes the choices made by the teams: 14 out of 24 teams used the video, which confirms the interest and potential of this new data modality. Only three teams used all the data at their disposal, but 15 teams used more than one data source. Another relevant point is that five teams further enriched the dataset with ancillary data (e.g., building footprints from OpenStreetMap, road maps from Google Street View, digital elevation models, and areal statistics). Indeed, this choice reminded that data of great quality are available at no cost on the Internet and can be included in processing pipelines to either make them more effective or decrease their computational load.

B. Winners

The 24 submissions were evaluated by the Award Committee and four papers were awarded. They were presented during the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) in Beijing, China. The papers awarded were as follows.
This section discusses the approaches proposed by the winners of the Contest. The work focused on a novel framework for spatiotemporal analysis of spaceborne remote sensing video data. For this purpose, a deep neural network was proposed to make use of high resolution satellite imagery in order to achieve a fine-resolution spatial scene labeling map. Moreover, a sophisticated approach was proposed to analyze activities and estimate traffic density from 150,000+ tracklets produced by a Kanade–Lucas–Tomasi (KLT) keypoint tracker. Both visual and quantitative analysis of the experimental results demonstrated the effectiveness of the proposed approach.

### A. Motivation

In comparison with the widely used static satellite imagery or video data acquired from unmanned aerial vehicles, space videos have remarkable advantages. For instance, a space video gathered from the ISS or by small satellites provides an opportunity to observe the dynamic states at urban or suburban scale and usually covers large areas. However, due to the relatively coarse spatial resolution, the poor appearance of moving targets (e.g., cars) creates ambiguities in tracking and challenges scene interpretation tasks, such as activity analysis and density estimation. In this paper, the focus was the semantic scene interpretation of space videos, using spatiotemporal analysis. More specifically, the aim was to answer the following three questions.

1) What land-use categories are observed in the video and where do they appear (spatial scene labeling)?
2) Which activities can be observed in the video (temporal activity analysis)?
3) And how dense is the traffic flow (traffic density estimation)?
Owing to their rich temporal information, video data can be exploited for activity analysis and traffic density estimation, which are classical computer vision applications. Since videos from space have only recently become available, their use to augment the semantic interpretation of covered geographic areas has rarely been addressed in the remote sensing community so far. In a pioneering work moving from static image analysis to video parsing, Kopsiaftis and Karantzalos [28] developed an automatic vehicle detection and traffic density estimation model for VHR satellite video data. This model was based on background estimation followed by background subtraction in every video frame. However, it could not obtain the trajectories of moving objects for further analysis of temporal activities. In addition, it was almost impossible to automatically obtain a stationary background model, since the camera was moving during the video acquisition (see Fig. 8). In this paper, the goal was to develop a sophisticated spatio-temporal scene analysis approach for space videos that was for the first time capable of performing spatial scene labeling and temporal activity analysis on multimodal image data.

In the following, the proposed approach will be summarized (see the processing chain in Fig. 7): first, a deep neural network architecture tailored to pixel-wise spatial scene labeling was proposed for spatial analysis (see Section IV-B); then, a tracklet-based temporal parsing method was introduced for activity analysis and traffic density estimation (see Section IV-C).

B. Spatial Scene Labeling via Deep Neural Network

For spatial scene labeling, the recent study [38] on the use of typical CNNs such as VGG16 [44] was extended to classify category-independent regions sampled from VHR images. In particular, unpooling and oversegmentation were exploited to refine the typically low resolution (LR) results produced by the CNN.

1) CNN: A typical convolutional network interleaves convolutional layers and pooling layers often followed by one or more fully connected layers. The specific architecture of the network was topologically identical to VGG16 [44].

The VGG16 network makes use of a stack of convolutional layers with a very small receptive field of $3 \times 3$, rather than using larger ones, such as $5 \times 5$ or $7 \times 7$. The reason is that $3 \times 3$ convolutional filters are the smallest kernels to capture patterns in different directions, such as center, up/down, and left/right, and bring multiple advantages. First, the use of stacks of small convolutional filters separated by rectified linear units (ReLUs) increases the nonlinearities inside the network. Then, the number of parameters to be learned decreases. The convolution stride in the network is fixed to 1 pixel. The spatial padding of the convolutional layer input is such that the spatial resolution is preserved after convolution, i.e., 1 pixel for the $3 \times 3$ convolutional layers. Spatial pooling is achieved by carrying out several max pooling layers, which follow some of the convolutional layers. Max pooling is performed over $2 \times 2$ pixel windows with stride 2. After several convolutional and pooling layers, the high-level reasoning in the neural network is achieved via fully connected layers.
2) Fine Tuning: Here, a fully connected layer was added and the new network was fine-tuned using the patches sampled from the multispectral image, which made the network significantly easier to train than a CNN trained from scratch. Fine tuning was performed by retraining the final fully connected layer. Fine tuning the ImageNet-trained VGG16 model for the specific task addressed here was clearly advantageous and yielded decent results, since the amount of labeled training data for the Vancouver images was insufficient. Finally, the fine-tuned network was applied to the first video frame for spatial scene labeling.

3) Refinement: Using pretrained architectures was very well suited at extracting features from data by spatially shrinking the feature maps. Pooling was necessary to allow gradually gathering information over the feature maps, and to make the network computationally feasible. However, this produced a scene labeling map with a reduced spatial resolution. To obtain a high-resolution fine-edged scene labeling map, it was necessary to refine the LR probability maps produced by the CNN. Several strategies have been proposed to perform this step, ranging from excluding spatial pooling in convolutional layers [45] to learning a deconvolution network to upsample the maps [14]. A simpler approach, shown in Fig. 7, was applied. The main ingredients were unpooling and region refinement.

Unpooling extended the feature maps, as opposed to pooling (see Fig. 10), and took the LR probability maps as input to produce VHR but grainy maps called coarse probability maps. Then, these coarse maps were projected to superpixels\(^3\) by assigning, to each superpixel, the average value of its corresponding region in the coarse probability maps as

\[
p(s_i) = \frac{1}{|s_i|} \sum_{c \in s_i} p(c) \tag{1}
\]

where \(s_i\) represents the \(i\)th superpixel, \(|s_i|\) is the number of pixels belonging to \(s_i\), \(c\) stands for a generic pixel in the superpixel, and \(p(\cdot)\) denotes probability map entries. In this way, the refined fine-edged probability maps were obtained. Fig. 9 compares the original LR output (top) with the refined one obtained by the proposed approach (bottom).

C. Temporal Analysis Using Tracklets

Object tracking for temporal analysis has been well studied in computer vision for more than two decades. Video from space, however, brings new methodological challenges, such as 1) the low frame rate, which undermines the common motion smoothness assumption; and 2) the limited spatial resolution of moving objects, which prevents the use of robust and discriminative appearance-based tracking algorithms.

1) Extracting Valid Tracklets: In this paper, tracklets, instead of the trajectories obtained by the object tracking algorithms, were used to analyze the activity. A tracklet is a fragment of a trajectory obtained by a tracker with a short period. Compared to long-term object tracking, they are more reliable. In our approach, a KLT keypoint tracker [47] was used to extract \(150000+\) tracklets from the space video [see Fig. 11(a)]. If the camera was stationary, the tracklets would have been entirely caused by real object movements (positive tracklets). Unfortunately, the camera kept slightly moving throughout the entire video acquisition (see the effect of such movements over \(30\) s in Fig. 4), which meant that the obtained tracklets included many negative samples caused by parallax. Therefore, it became necessary to remove negative tracklets before carrying out any follow-up activity analysis. Here, a prior constraint on the tracklet denoising was enforced according to the observation that the displacements of moving object tracklets were usually larger than the displacements of still object tracklets (e.g., buildings) caused by the camera movement. In this paper, the threshold value was fixed to \(11\) pixels, which corresponded to a minimum car speed of \(20\) km/h, given the spatial resolution of the video.

\(^3\)A superpixel can be defined as a set of locally connected similar pixels that preserve detailed edge structures for a fine segmentation. The code used for generating superpixels is available at http://coewww.rutgers.edu/riul/research/code/EDISON/ [46]. In this paper, spatial and range bandwidth parameters were set to \(7\) and \(6.5\), respectively.
Regarding activity analysis, the standard quantitative measure of precision is based on the number of true positives (TP) and false positives (FP) and is computed as precision = TP / (TP + FP). For quantitative evaluation, the whole visible area of the video was used and tracklet ground truth was manually determined for 30 s. The precision of the temporal activity result reached 95.3%. Fig. 13 provides a close zoom of the activity analysis in which two moving cars were detected and their moving directions and speeds could be estimated. In addition, the traffic density estimation result is presented in the bottom right image of Fig. 12. The estimated density highly correlates with the actual traffic situations, e.g., high values in main roads and intersections.

It is also worth noting that, here, the focus was on addressing two relatively independent problems, i.e., spatial scene labeling and temporal activity analysis. Accordingly, an “open-loop” solution in which the scene labeling was accomplished without making use of the temporal analysis results was designed. Indeed, a “closed-loop” system, in which the spatial scene labeling is further refined using the outcome of the temporal analysis, will represent an interesting future extension of this work.

V. SIMULTANEOUS REGISTRATION, SEMANTIC SEGMENTATION, AND CHANGE DETECTION FROM MULTISENSOR, MULTITEMPORAL SATELLITE IMAGE PAIRS

This section details the method proposed by the second ranked team of the Contest.

A. Motivation

The current generation of space-borne and airborne imaging sensors are generating nearly continuous streams of massive, multitemporal, high-resolution remote sensing data. However, in order to efficiently exploit these datasets, their accurate coregistration is the first indispensable processing step along with any further analysis and change detection procedures. In the framework of the DFC 2016, a methodology was proposed that was able to ingest information from multisensor and multitemporal satellite image pairs, by addressing concurrently the tasks of registration, semantic segmentation, and change detection. These three tasks are usually addressed separately.

A number of studies were focused on jointly tackling the problem of registration and semantic segmentation for mainly video sequences or medical images [48]–[51]. Similar research efforts were focused on jointly addressing the tasks of segmentation and tracking in image video sequences [52]–[54]. However, such formulations cannot exploit sparse multitemporal datasets with changes in-between the various acquisition dates. To tackle such problems, a framework was designed for jointly register images, detect changes, and producing semantic segmentation maps at the same time.

In particular, the formulation of [6] and [55], was extended by adding a graph tackling the semantic segmentation problem. The developed method jointly integrated the following three types of energies into an energy minimization framework:
data-driven costs issued from classification scores\footnote{Throughout Section V, we will use the word \textit{classification} for the scores issued from this independent classification and \textit{semantic segmentation} for those obtained by the proposed energy minimization method.} (various classification algorithms can be employed, e.g., [56]);

2) registration metrics (e.g. similarity metrics);

3) change detection scores.

These energies were efficiently coupled with local geometric constraints in the context of a higher order graph. Reduction methods were used to map this graph into a pairwise one, which was optimized using efficient linear programming. The experimental results were very promising: less than 2 pixels in terms of mean displacement error for the registration; above 77\% in most cases regarding the completeness and correctness rates of the semantic segmentation; and change detection accuracy higher than 70\%.

\subsection{Methodology}

\subsubsection{Graph Formulation:} Without loss of generality, three distinct graphs with exactly the same topology/geometry and of lower resolution than the image dimensions were considered. They corresponded to a lower resolution grid than the input image and every original pixel was assigned to the closest graph node to compute the related energy contributions. Following the notations of [6] and [55], the first graph, $G_{\text{reg}}$, involved nodes where the labels corresponded to deformation vectors from the registration process, i.e., a mapping between the source and the target images. The second graph, $G_{\text{ch}}$, referred to nodes with binary labels expressing changes in the temporal domain. The third graph, $G_{\text{seg}}$, which was the one introduced in this work, referred to the labels representing the semantic segmentation of the image. In terms of connectivity, each graph was endowed with a first-order neighborhood system accounting for local consistency on the corresponding label space. Furthermore, two intergraph connections were considered, one associating the corresponding nodes in the detection and registration graphs, and the other involving hypercliques that linked the corresponding nodes of all graphs.

This graph structure was superimposed on the image in such a way that each node of the graph depended on a subset of pixels in its vicinity. The objective function was defined on the entire image domain, and every pixel contributed to the graph with a weight depending on its distance from a node in the support domain. The graph structure defined the support domain and the computational complexity of the implementation, thus, a finer grid generated a higher computational load. In particular, the dimensions of the graph were related to the image size, thus representing a tradeoff between accuracy and computational complexity.

\subsubsection{Energy Formulation:} By integrating the three graphs in a single framework, the proposed energy was a function $E_{\text{reg, ch, seg}}(l^{\text{reg}}, l^{\text{ch}}, l^{\text{seg}})$, in which the labels of each node $p$ of the coupling graph were $l_{p} = [l^{\text{ch}}_{p}, l^{\text{seg}}_{p}, l^{\text{reg}}_{p}]$. The term $l^{\text{ch}}_{p} \in \{0, 1\}$ represents the label for change detection, $l^{\text{seg}}_{p} \in \{0, 1\}$ represents the label for binary semantic segmentation, and $l^{\text{reg}}_{p} \in \Delta$ is the registration label. $\Delta = \{d^{1}, \ldots, d^{n}\}$ corresponds to all possible displacements. The label space could be described as $L = \{0, 1\} \times \{0, 1\} \times \Delta$. In the next section, we detail the key ideas of all the terms used to calculate the energy function.

\subsection{Energy Terms}

The energy was composed of a series of terms related to the three tasks of change detection, registration, and semantic segmentation. The energy terms could be either singletons (i.e., functions of the score of one specific node), pairwise (i.e., depending on the spatial neighborhood for one single task), or coupled (i.e., using the results of two or more tasks on a single node simultaneously).

\subsubsection{Singleton and Pairwise Terms:} Both the registration and change detection terms depended on the two considered images simultaneously, and therefore, only had pairwise energy contributions. Regarding the semantic segmentation process, the goal
was to assign the correct semantic segmentation label to each node of the target image. In particular:

1) the registration \( V_{pq, \text{reg}}(l_p^{\text{reg}}, l_q^{\text{reg}}) \) and change detection \( V_{pq, \text{ch}}(l_p^{\text{ch}}, l_q^{\text{ch}}) \) pairwise terms followed the same formulation as in [6] and [55] and penalized neighboring nodes with different registration or change detection labels;

2) the semantic segmentation graph contained a term with the classification score for each node (i.e., a singleton term) \( V_{\text{seg}}(l_p^{\text{seg}}) \) and a pairwise term \( V_{\text{seg,seg}}(l_p^{\text{reg}}, l_q^{\text{seg}}) \), which penalized different semantic segmentation labels in neighboring nodes. The two terms followed the same formulation as in [40].

2) Coupled Intergraph Energy Terms: As stated previously, two intergraph connections were considered. They translate into two coupled terms addressing two and three tasks simultaneously.

1) The coupling between the registration and change detection terms \( V_{\text{reg, ch}}(l_p^{\text{reg}}, l_p^{\text{ch}}) \) was achieved through the interconnection between the two graphs. In the absence of change, the cost was calculated using a similarity function, while in the presence of change, a fixed cost was used.

2) The coupling between the three terms was performed using one potential term \( V_{\text{reg, ch, seg}}(l_p^{\text{reg}}, l_p^{\text{ch}}, l_p^{\text{seg}}) \) that penalized different semantic segmentation labels in the source and target images for all possible displacements in cases of no-change and the same ones in cases of change.

3) Final Global Energy: The global energy function was a linear combination of the aforementioned energy terms, weighted by six nonnegative weight parameters \( w_1, w_2, \ldots, w_6 \). These parameters should be optimized prior to the minimization of the energy with respect to the labels \( l_p^{\text{ch}}, l_p^{\text{reg}}, \) and \( l_p^{\text{seg}} \) of all nodes \( p \). Details on the specific formulation of the energy can be found in [40].

D. Experimental Results and Evaluation

The evaluation of the developed framework was performed on all the data described in Section II. From now on, we will refer to the first Deimos-2 acquisition (March 2015) as \( D1 \), to the second Deimos-2 image (May 2015) as \( D2 \), and to the Iris video as \( V \). The level 1C image products were used, and were further radiometrically corrected, and then, pan-sharpened based on the standard high-pass filter method, resulting into an overlapping image pair of approximately \( 12760 \times 11000 \) pixels. The area of the overlap between the image pairs and the Iris video frames was approximately \( 4720 \times 2680 \) pixels. In order to employ an additional image/map, which could serve as a reference/target map while contributing to the automation of the subsequent training procedure, an image mosaic \( G \) and the corresponding land map\(^6\) were downloaded from Google Maps APIs.\(^7\)

---

\(^6\)By “land map” we intend the raster simplified view that can be used in Google Maps as an alternative to the satellite image.

\(^7\)All the considered raw unregistered data and ground truth data, and several experimental results can be viewed at: http://users.ntua.gr/karank/Demos/DemoContest16.html

---

### Algorithm 1: General formulation of the algorithm proposed by the second ranked team.

Required: param: parameters of the algorithm

Ensure: Registration, Semantic Segmentation, and Detection labels.

1: \( i \leftarrow \text{param.mingridlevel} \)
2: while \( i++ < \text{param.maxgridlevel} \) do
3: \( G \leftarrow \text{create graph} \)
4: \( U \leftarrow \text{unary potentials} \)
5: \( B \leftarrow \text{pairwise potentials} \)
6: \( \eta() \leftarrow \text{projection function (see [40])} \)
7: for all node \( p \in G \) do
8: for all pixel \( j \) in the vicinity of \( p \) do
9: \( z_{pj} \leftarrow \eta(j) \)
10: end for
11: end for
12: for all label \( l_p \in \{0, 1\} \times \{0, 1\} \times \Delta \) do
13: for all pixel \( j \) do
14: \( \sum_j = w_1 V_{\text{reg, ch}}(l_p) + w_3 V_{\text{seg}}(l_p) + \)
15: \( w_2 V_{\text{reg, ch, seg}}(l_p) \)
16: end for
17: end for
18: for all label \( l_p \) do
19: for all label \( l_q \) do
20: \( B_{pq}(l_p, l_q) = w_4 V_{pq, \text{reg}}(l_p, l_q) \)
21: end for
22: end for
23: \( l \leftarrow \text{FastPD-Optimization}(U, B) \)
24: for all node \( p \) do
25: for all pixels \( j \) in the vicinity \( p \) do
26: \( l_j = z_{pj} \cdot l_p \)
27: end for
28: end for
29: Transform the source image
30: end while

---

1) Implementation Details: The parameters of the developed framework were not constant across different datasets. Grid search over a small area of the dataset was used for fine tuning all the parameters. Then, this set of parameter values was applied to the entire dataset. This was of course a limitation of the developed method, however, with satellite or aerial data of similar spatial and spectral resolution, significant differences were not expected. Details on techniques and implementation can be found in [40], while for the optimization procedure, FastPD was employed [57]. For an overview of the method, see the pseudocode in Algorithm 1.

2) Classification Scores: In order to estimate the classification scores, a patch-based deep learning framework was used. It reported high classification accuracy rates for a number of cases [56], [58], [59]. These deep architectures require a large number of training data. To this end, the Roads, Buildings, Sea, and Vegetation classes were automatically extracted from the
Google land map (based on the different colors). To do so, all datasets were registered to the Google image mosaic in order to relate every pixel to one of the Google land map colors. Furthermore, polygons for the Vessels, Building shadow, Vegetation shadow, Soil, and Clouds classes were manually annotated. For the Iris video sequence, Vegetation shadows was merged with Vegetation. Spectral analysis on the derived numerous polygons and probabilities was employed (as in [58]) in order to create the final training and validation sets for the eight terrain classes, i.e., Roads, Buildings, Building Shadows, Soil, Sea, Ship/vessels, Vegetation, and Vegetation shadows.

The training for the experiment considering the D1 and D2 images was performed on the large (12760 × 11000 pixels) overlapping region with eight classes. On the contrary, seven classes were employed for V. Based on the derived polygons, numerous patches of size 21 × 21 including all the spectral bands were created by centering each patch on the annotated pixels [59]. Approximately 200 000 randomly selected patches per class were used for the Deimos-2 data, while 50 000 were used for the Iris video. A CNN with ten layers was trained: the network was made of two convolutional layers, each followed by tangent and max pooling layers, and ended with two fully connected, a tangent, and a linear layer. The model was trained with a learning rate equal to 1 for 40 epochs; afterwards, the learning rate was halved every 2 epochs. The same CNN architecture and configuration settings were used for the classification of the Iris video frames.

3) Registration Results: For the validation of the registration results (see Fig. 14), several ground control points were manually collected in all resulting image pairs. In Table I, the mean displacement errors for both axis (Dx and Dy) and the distance (DS) in pixels are presented. It should be mentioned that the registration process did not manage to address the largest relief displacements of the tallest buildings/skyscrapers of this part of Vancouver, and these errors hindered the results of both semantic segmentation and change detection. All other building rooftops, roads, and terrain classes were registered with subpixel accuracy. A quantitative comparison is also provided in Table I: These results are compared with those obtained by the algorithm in [60] for both tasks of registering the different sources to the Google mosaic and registering each frame of the video sequence to the first frame. One can observe that both methods resulted in subpixel accuracy, indicating that the coupling of registration priors with semantic segmentation and change detection priors does not harm the quality of the registration.

4) Semantic Segmentation Results: The framework was validated for the detection of three different classes out of the original eight, namely: Buildings, Ship/vessels, and Vegetation. For the semantic segmentation and change detection results, the quantitative evaluation of the framework was performed using the completeness, correctness, and overall quality criteria at the object level. In particular, after an intensive manual
photointerpretation procedure, ground-truth/reference data were created for the three aforementioned classes in four different smaller regions. The resulting true positives, false negatives, and false positives were calculated on the validation dataset after the application of the developed framework. The overall accuracy (OA) and average accuracy (AA) were also calculated. After the optimization and based on the polygons of the Sea class, which were derived automatically from Google Maps, all objects segmented as Buildings in the Sea and as Ship/vessels in the land areas were neglected.

As one can observe in Fig. 15, although the original classification scores constrained significantly the result, the developed framework could improve the semantic segmentation results in several image regions as compared with the classification accuracies obtained by the CNN. In particular, the quantitative evaluation (see Table II) indicated that the detection completeness rates were above 78% (apart from the Buildings class in the D1–D2 pair), and the detection correctness rates were above 72% in all cases. The highest rates were for the Vegetation class, indicating that the NIR Deimos-2 band significantly contributed to class separation. Most semantic segmentation errors were due to false alarms near the port, pier, and ship wake on the sea. Moreover, the Buildings and Roads classes were confused in certain cases.

5) Change Detection Results: Similar qualitative errors were observed on the change detection results (see Fig. 16) obtained from both image pairs. Quantitative results (see Table III) suggested lower completeness and correctness rates as compared to the semantic segmentation task, as expected. These results were mainly due to a number of false positives

### Table III

Quantitative Evaluation of the Change Detection Results (Per Class Results from [40])

<table>
<thead>
<tr>
<th></th>
<th>Completeness</th>
<th>Correctness</th>
<th>Overall Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ship/vessel</strong></td>
<td>68.6%</td>
<td>66.7%</td>
<td>66.2%</td>
</tr>
<tr>
<td><strong>Vegetation</strong></td>
<td>88.2%</td>
<td>82.3%</td>
<td>75.6%</td>
</tr>
<tr>
<td><strong>Buildings</strong></td>
<td>69.2%</td>
<td>67.4%</td>
<td>57.4%</td>
</tr>
<tr>
<td><strong>Overall (proposed)</strong></td>
<td>75.3%</td>
<td>72.1%</td>
<td>66.4%</td>
</tr>
<tr>
<td><strong>OA, method of [6]</strong></td>
<td>70.3%</td>
<td>67.1%</td>
<td>52.3%</td>
</tr>
<tr>
<td><strong>Iris July [V]—Deimos-2 May [D2]</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Ship/vessel</strong></td>
<td>70.6%</td>
<td>69.5%</td>
<td>53.8%</td>
</tr>
<tr>
<td><strong>Vegetation</strong></td>
<td>81.1%</td>
<td>79.6%</td>
<td>67.2%</td>
</tr>
<tr>
<td><strong>Buildings</strong></td>
<td>71.3%</td>
<td>65.6%</td>
<td>51.9%</td>
</tr>
<tr>
<td><strong>Overall (proposed)</strong></td>
<td>74.3%</td>
<td>71.6%</td>
<td>57.6%</td>
</tr>
<tr>
<td><strong>OA, method of [6]</strong></td>
<td>70.4%</td>
<td>66.3%</td>
<td>51.8%</td>
</tr>
</tbody>
</table>

---

in the dense urban regions where the relief displacements were significant due to the tallest buildings and skyscrapers.

Additionally, experiments with the unsupervised change detection method in [6] were also performed to compare the performance with and without semantic segmentation labels. The proposed method not only provided additional information (i.e., semantic segmentation maps and from-to change trajectory maps) but also resulted in higher accuracy rates. In particular, the developed system led to an improvement of at least 6% in the mean overall quality compared to the method in [6], indicating that the use of the semantic segmentation labels allows the number of false alarms to be reduced and more accurate change maps to be produced.

Finally, the results of the proposed method were compared with those of [6] and [60] in terms of computational performance. For the proposed higher order formulation, reduction methods were employed to reformalize as a pairwise model, which is indeed quite computationally demanding. In particular, for a $1000 \times 1000$ pixel image, the deformable registration component alone (see [60]) required 2–3 min. Then, the registration and the unsupervised change detection (like in [6]) converged after 6 min. The proposed framework converged after approximately 12 min with the same configuration setting.

VI. CONCLUSION

In this paper, we discussed the scientific results of the 2016 IEEE GRSS DFC organized by the IEEE GRSS Image Analysis and Data Fusion Technical Committee. We described the dataset and described the overall scientific outcomes of the competition, by first presenting its overall results, and then, focusing on the strategies proposed by the first and second place teams. These teams made use of all the data available and developed methodologies rooted in the latest advances in computer vision and machine learning: The winners proposed a CNN and a tracking algorithm, while the second place team developed a random field model performing multiple tasks simultaneously. Both methodologies were effective and tailored to the peculiarities of the new generation remote sensing data that were released for the contest, including in particular video from space (introduced for the first time in this annual competition).

Moreover, the results of the DFC 2016 also exceeded these two individual methods: The organizers appreciated the originality of all the submissions and the hard work of all the participating teams. The participants’ efforts in stretching the current limits of the discipline were remarkable, both methodologically—by integrating advanced video processing and computer vision techniques—and application wise. It also was highly appreciated that new challenging fields of study, such as population density estimation (proposed by the third place team [41]) and human activity modeling (from the winning team), were addressed within the contest.

The data will remain downloadable for free from the IEEE GRSS website. Ground references were made available by the two winning teams (see Sections IV and V for the URLs). We do hope that these data will serve to push the boundaries of remote sensing data fusion even further and make video form space a new valuable data source for remote sensing applications.

ACKNOWLEDGMENT

The authors would like to express their greatest appreciation to Deimos Imaging and Urthecast, for acquiring and providing the data used in the competition and for indispensable contribution to the organization of the Contest, and the IEEE GRSS for continuously supporting the annual Data Fusion Contest through funding and resources.

REFERENCES
