report

by Ucd Ucd

Submission date: 02-Nov-2022 10:23PM (UTC+0530)

Submission ID: 1942553511

File name: WP3_D3.1_report.docx (245.15K)

Word count: 3474

Character count: 19177

DTIF3 Contract Reference: DT 2020 0209

Title/Acronym: **C**reating an **A**rchitecture for **M**anipulating **E**arth **O**bservation data (CAMEO)

WP3

(Data Quality Assurance)

Project Workplan, Deliverables:

GANTT Chart: Timing of Work Packages and their components

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18
WP0	D0.1	D0.2										D0.3						
WP1								D1.1			D1.2	D1.3.1						D1.3.2
WP2								D2.1		D2.2	D2.3						D2.5.1	D2.2
WP3																		
WP4									D4.1			D4.2						
WP5																		D5.1
WP6			D6.1, D6.2									D6.3						D6.4
WP7					D7.1			D7.2										

	M19	M20	M21	M22	M23	M24	M25	M26	M27	M28	M29	M30	M31	M32	M33	M34	M35	M36
WP0						D0.4												D0.5
WP1									D1.3.3									
WP2	D2.4	D2.3						D2.2						D2.5.2			D2.3	
WP3	D3.1					D3.2												
WP4		D4.3						D4.4										
WP5		D5.3, D5.4			D5.2		D5.3					D5.3		D5.3				
WP6			D6.6.1									D6.6.2				D6.5	D6.6.3	
WP7				D7.3.1		D7.4.1				D7.3.2		D7.4.2	D7.3.3		D7.4.3			D7.5

Work package number	WP3		**Start Date**			M6	
Work package title	Data Quality Assurance						
Participant number	1	2	3	4	5	6	7
Short name of participant	**UCD**	VC	ES	ICON	TM	TWM	Dell Technologies
Person/months	**65 (-5)***	4	4	3	16	8	6

Objectives

●Design and formulation of mechanisms for the adjudication of data quality;

●Use of discovery services to identify temporally and geographically adjacent data sources;

●Provision of services for ground truthing data with relevant (location and temporal adjacency) and known high quality data sets;

●Design and implementation of trusted mechanisms to filter 'poor quality' data and ensure non-admittance to the data warehouse

Description of work

Poor quality data will invariably lead to poor decisions. It is imperative therefore to seek to ensure that the CAMEO data warehouse is only populated with quality data or at the very least data for which the indicative quality of data is known.
Adjudication of data quality and mechanisms for doing so need to be incorporated throughout the entirety of the big data model including data collection, data pre-processing, data processing and analytics, and data use. This work package will involve 4 subtasks.

Deliverables (brief description and month of delivery)

D3.1 Design of Data Quality Adjudication Framework (M19)

D3.2 Design and Implementation of Data Quality Filter (M24)

Milestones

MS3.1 Delivery of Data Quality Filter (M24)

Deliverables for D3.1

1. **Categorization of earth observatory data types**
2. **Identifying need of Data Quality and its evaluation matrix**
3. **Implementation of EO raster data Quality matrix**
4. **Predictive analysis of usability for a SME**

Deliverables for D3.2

1. **Classification of various vector EO data and identification of Vector data quality metrics**
2. **Implementation of EO vector data Quality matrix**
3. **Predictive analysis of usability of vector data for SME**

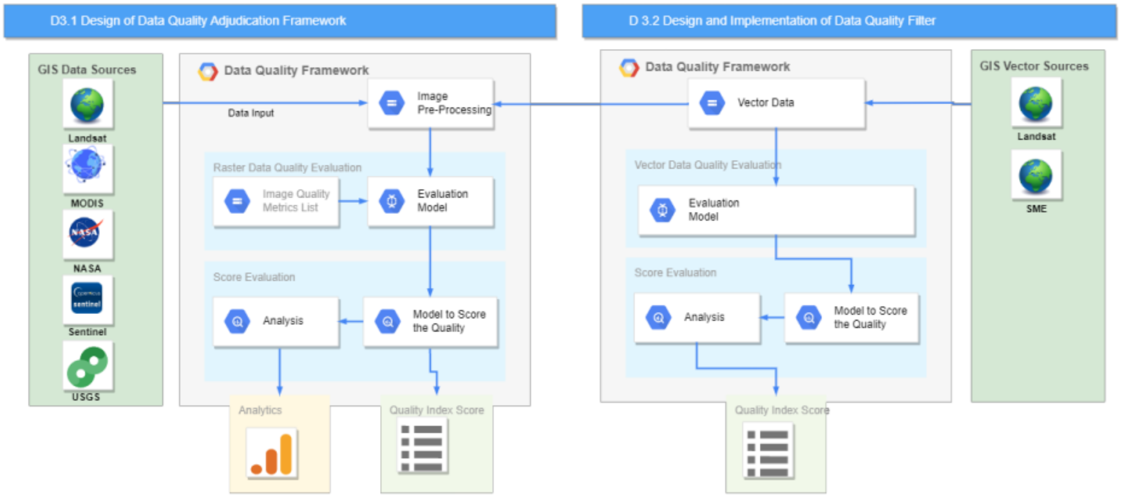


Figure 1. WP3 Architecture

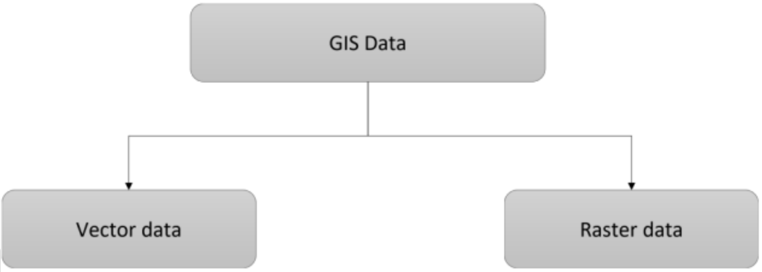
Figure 1 shows the architecture and working of work package 3 where the WP3 is distributed into two deliverable s based on the data quality evaluation of Image data and vector data. The first portion D3.1 of the architecture is collection and evaluation of Satellite image dataset based on various data quality metrics. The module is responsible for evaluating a score out of the various indexes that will be evaluated. The score will define the quality of GIS data. Similarly in D3.2 the module is defined to collect vector data in form of GIS layers and checking for data quality and error in the data based on comparison from existing models. The outcome of the package will be an analysis to the used about the quality and usability of the GIS data based of the use cases of GIS data in real-world.

Task 3.1: Design and formulation of mechanisms for the adjudication of data quality;

Abstract:

A series of data quality services will be developed the first tranche of which will focus upon the quality of collected data. In order to assist with such a series of services will be developed by which to identify and source relevant (location and temporal adjacency) data of known high quality which can be used to affirm data quality and support ground truthing. UCD has established experience in data quality research [3,4,5] examining data quality in terms of data trust. UCD has also established credentials as part of AIREO (AI-Ready Earth Observation Training Datasets) project exploring automated quality assessment together with best-practices around dataset documentation in relation to quality - provenance information, and information pertaining to data collection protocols (e.g. including for non-EO derived ground truth/reference data/annotations/labels). A data quality coefficient will be determined that will somewhat crudely apportion a measure to data. Quality will be assessed across numerous dimensions including completeness, adjacency (spatial & temporal), lossiness, noise but also other factors including suitability for ML use cases e.g. are the labels/annotations of suitable volume, quality, and class distributions. Subsequent quality service bundles will address quality measures across the big data model stages; pre-processing, conflation, analytics and usage.

GIS Data Types [6]



There are two different types of GIS data, vector data and raster data. Each type of data has its own format.

Vector Data [7]

Vector data is the spatial data most people are familiar with, as it is the format presented in mapping portals such as Open Street Maps and Google Maps. It is also used extensively in computer graphics and computer-aided design (CAD). It consists of points, lines, and polygons.

Point Data – Point Data typically represents nonadjacent features or distinct data points. Points are zero dimension, so you cannot measure their length or area. Examples of point data would be cities, points of interest, and schools.

Line Data – Line data is also known as arc data. It represents linear features such as rivers, streets, and trails. Line data has a starting and an ending point, and, since it only has one dimension, can only be used to measure length.

To distinguish arc features from each other, some lines may be solid while others are dashed, and different colours or line thicknesses may be used. For example, a road may be a solid black line, while a river is a dashed blue line.

Polygon Data – Polygons typically represent areas such as cities, lakes, or forests. Unlike point and line data, polygons are two dimensional and can measure the perimeter or area of a geographic feature. Colour schemes, patterns or gradation colour schemes could be used to identify polygon features.

Vector images are high-quality representations of an image or a shape. They can be enlarged or reduced with no loss of quality. To create or manipulate a vector image, you must use a program like Adobe Illustrator. A camera cannot capture a vector image.

Raster Data

Raster data, also known as grid data, is made up of pixels, and each pixel has a value. You will typically find raster data on topographic maps, satellite images, and aerial surveys. Raster data is vital for meteorology, disaster management, and industries where analysing risk is essential.

There are two types of raster data, continuous and discrete.

Continuous Data – Continuous rasters are cells on the grid that gradually change. Some examples would be an aerial photo, elevation and temperature. Continuous raster surfaces come from a fixed registration point. For instance, in digital elevation models, sea level is used as a registration point. Each cell represents a value that is above or below sea level.

Discrete Data – Discrete rasters have a specific theme or class, and each pixel is assigned to a specific class. Unlike continuous data, discrete data can only take specific values, not values within a range. For example, in a discrete raster land cover/use map, you can see each thematic class, and where it begins and ends is defined.

Unlike vector data, raster data is not scalable. If it is enlarged too much, it will get pixelated, and if stretched too much, it will become distorted. A digital photo is an example of raster data.

Raster data mainly covers satellite images from various sources like drone, satellite, heat maps and many more. Raster image file types include BMP, TIFF, GIF, and JPEG.

Spatial data Quality Components [1,2,3]

Data quality is the degree of data excellency that satisfy the given objective. In other words, completeness of attributes in order to achieve the given task can be termed as Data Quality. Production of data by private sector as well as by various mapping agencies assesses the data quality standards in order to produce better results. Data created from different channels with different techniques can have discrepancies in terms of resolution, orientation and displacements. Data quality is a pillar in any GIS implementation and application as reliable data are indispensable to allow the user obtaining meaningful results.

Spatial Data quality can be categorized into Data completeness, Data Precision, Data accuracy and Data Consistency.

- Data Completeness: It is basically the measure of totality of features. A data set with minimal amount of missing features can be termed as Complete-Data.
- Data Precision: Precision can be termed as the degree of details that are displayed on a uniform space. More about precision: GIS Data: A Look at Accuracy, Precision, and Types of Errors
- Data Accuracy: This can be termed as the discrepancy between the actual attributes value and coded attribute value.
- Data Consistency: Data consistency can be termed as the absence of conflicts in a particular database.

Motivation

EO data quality is managed at multiple levels by different partners. It may be good to know which levels matter to the respondents, what information about data quality is available/relevant/important, and who is in charge of data quality assurance at the relevant levels. In Ireland, the EPA coordinates national teams to validate information products from the Copernicus Land Monitoring Services (CORINE landcover, Forest data series, Water and Wetness, Natura - information on hotspots for nature conservation). In some cases, the work leads to correcting the data from Copernicus by integrating in-situ measurements and local information. In those cases, the EPA maintains a verified/corrected version of the data and provided it back to Copernicus. In other cases (e.g. Water and Wetness), verification shows that the data quality is insufficient but no correction is known. It would be good to identify similar work done nationally. It may also be good to know if the SMEs benefit from the correction done by the EPA. (Some SMEs indicate that they are using Copernicus products provided by the EPA).

Some SMEs are using Landsat data. The Landsat archive went through two significant reprocessing rounds to specifically improve data quality. Those result in two different Landsat data "Collection" - named Collection 1 and Collection 2. It may be good to see if the SMEs use data from which collection and if the reprocessing makes any difference to what they do.

Image Quality Metrix

metric	class	description	better	range	ref
Peak signal-to-noise ratio (PSNR)	FR	The ratio of the maximum pixel intensity to the power of the distortion.	higher	[0, inf)	[WIKI]
Structural similarity (SSIM) index	FR	Local similarity of luminance, contrast and structure of two image.	higher	(?, 1]	[paper] [WIKI]
Multi-scale structural similarity (MS-SSIM) index	FR	Based on SSIM; combine luminance information at the highest resolution level with structure and contrast information at several down-sampled resolutions, or scales.	higher	(?, 1]	[paper] [code]
Learned perceptual image patch similarity (LPIPS)	FR	Obtain L2 distance between AlexNet/SqueezeNet/VGG activations of reference and distorted images; train a predictor to learn the mapping from the distance to similarity score. Trainable.	lower	[0, ?)	[paper] [official repo]
Blind/referenceless image spatial quality evaluator (BRISQUE)	NR	Model Gaussian distributions of mean subtracted contrast normalized (MSCN) features; obtain 36-dim Gaussian parameters; train an SVM to learn the mapping from feature space to quality score.	lower	[0, ?)	[paper]
Natural image quality evaluator (NIQE)	NR	Mahalanobis distance between two multi-variate Gaussian models of 36-dim features from natural (training) and input sharp patches.	lower	[0, ?)	[paper]
Perception based image quality evaluator (PIQE)	NR	Similar to NIQE; block-wise. PIQE is less computationally efficient than NIQE, but it provides local measures of quality in addition to a global quality score.	lower	[0, 100]	[paper]

FR: Full Reference (USED AND CITED BY THE AUTHORS)

NR: No Reference (Not Popular)

Image Quality Metrics

1. **Peak signal-to-noise ratio (PSNR):**

Peak signal-to-noise ratio (PSNR) is an engineering term for the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. Because many signals have a very wide dynamic range, PSNR is usually expressed as a logarithmic quantity using the decibel scale.

2. Structural similarity (SSIM) index

   SSIM is a perception-based model that considers image degradation as perceived change in structural information, while also incorporating important perceptual phenomena, including both luminance masking and contrast masking terms. The difference with other techniques such as MSE or PSNR is that these approaches estimate absolute errors. Structural information is the idea that the pixels have strong inter-dependencies especially when they are spatially close. These dependencies carry important information about the structure of the objects in the visual scene. Luminance masking is a phenomenon whereby image distortions (in this context) tend to be less visible in bright regions, while contrast masking is a phenomenon whereby distortions become less visible where there is significant activity or "texture" in the image.

3. Multi-scale structural similarity (MS-SSIM) index [9]

   The structural similarity image quality paradigm is based on the assumption that the human visual system is highly adapted for extracting structural information from the scene, and therefore a measure of structural similarity can provide a good approximation to perceived image quality. This multiscale structural similarity method, which supplies more flexibility than previous single-scale methods in incorporating the variations of viewing conditions. We develop an image synthesis method to calibrate the parameters that define the relative importance of different scales. Experimental comparisons demonstrate the effectiveness of the proposed method.

4. Learned perceptual image patch similarity (LPIPS)[10]

   The Learned Perceptual Image Patch Similarity (LPIPS_) is used to judge the perceptual similarity between two images. LPIPS essentially computes the similarity between the activations of two image patches for some pre-defined network. This measure has been shown to match human persеption well. A low LPIPS score means that image patches are perceptual similar.

5. Gradient Magnitude Similarity Deviation (GMSD) [11]

The image gradients are sensitive to image distortions, while different local structures in a distorted image suffer different degrees of degradations. This motivates us to explore the use of global variation of gradient based local quality map for overall image quality prediction. We find that the pixel-wise gradient magnitude similarity (GMS) between the reference and distorted images combined with a novel pooling strategy the standard deviation of the GMS map can predict accurately perceptual image quality. The resulting GMSD algorithm is much faster than most state-of-the-art IQA methods, and delivers highly competitive prediction accuracy.

6. Feature Similarity Index Model (FSIM)

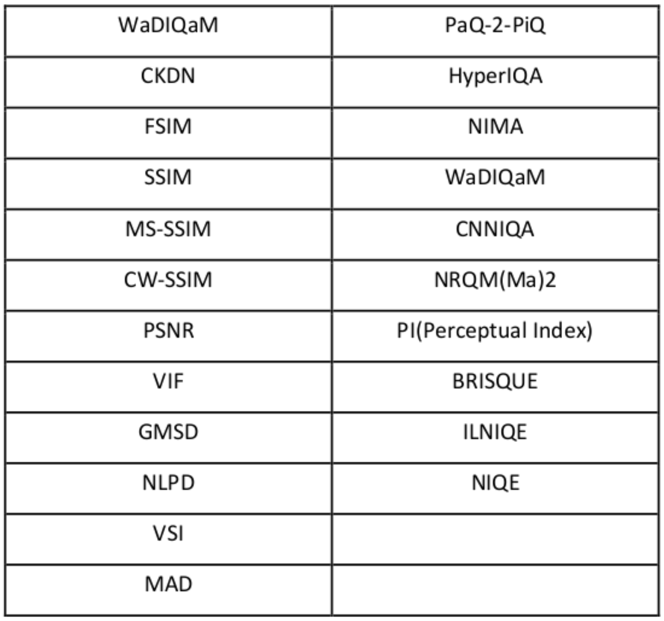
Image quality assessment (IQA) aims to use computational models to measure the image quality consistently with subjective evaluations. The well-known structural similarity index brings IQA from pixel- to structure-based stage. The feature similarity (FSIM) index for full reference IQA is proposed based on the fact that human visual system (HVS) understands an image mainly according to its low-level features. Specifically, the phase congruency (PC), which is a dimensionless measure of the significance of a local structure, is used as the primary feature in FSIM. Considering that PC is contrast invariant while the contrast information does affect HVS' perception of image quality, the image gradient magnitude (GM) is employed as the secondary feature in FSIM. PC and GM play complementary roles in characterizing the image local quality. After obtaining the local quality map, we use PC again as a weighting function to derive a single quality score. Extensive experiments performed on six benchmark IQA databases demonstrate that FSIM can achieve much higher consistency with the subjective evaluations than state-of-the-art IQA metrics.

7. Visual Saliency-based Index

Visual saliency (VS) has been widely studied by psychologists, neurobiologists, and computer scientists during the last decade to investigate, which areas of an image will attract the most attention of the human visual system. Intuitively, VS is closely related to IQA in that suprathreshold distortions can largely affect VS maps of images. With this consideration, we propose a simple but very effective full reference IQA method using VS. In our proposed IQA model, the role of VS is twofold. First, VS is used as a feature when computing the local quality map of the distorted image. Second, when pooling the quality score, VS is employed as a weighting function to reflect the importance of a local region. The proposed IQA index is called visual saliency-based index (VSI).

Categorization of Image quality metrics

FR Method	**NR Method**
AHIQ	FID
PieAPP	MANIQA
LPIPS	MUSIQ
DISTS	DBCNN



WaDIQaM	PaQ-2-PiQ
CKDN	HyperIQA
FSIM	NIMA
SSIM	WaDIQaM
MS-SSIM	CNNIQA
CW-SSIM	NRQM(Ma)2
PSNR	PI(Perceptual Index)
VIF	BRISQUE
GMSD	ILNIQE
NLPD	NIQE
VSI	
MAD	

Task 3.2: Design and Delivery of discovery services to identify temporally and geographically adjacent data sources.

Abstract

A data discovery service will be developed to identify adjacent data sources which will serve to underpin and inform determination of data quality coefficient(s). Data can be ground truthed through cross comparison of a given data stream with known data of high quality and adjacent within the spatio-temporal domain data. This discovery service will support identification of data sources access a wide range of data categories, IoT enabled devices, third party databases/sets, citizen derived data, satellite imagery and drone data.

Database:

- Landsat [8]

The Landsat program is the longest-running enterprise for acquisition of satellite imagery of Earth. It is a joint NASA / USGS program. Landsat shows us Earth from space. Since the first Landsat satellite launched in 1972, the mission has collected data on the forests, farms, urban areas and freshwater of our home planet, generating the longest continuous record of its kind. Decision makers from across the globe use freely available Landsat data to better understand environmental change, manage agricultural practices, allocate scarce water resources, respond to natural disasters and more.

With the launch of Landsat 9 scheduled for mid-2021, the mission will continue its legacy of monitoring key natural and economic resources from orbit. Landsat 9, managed by NASA's Goddard Space Flight Center in Greenbelt, Maryland, will carry two instruments: the Operational Land Imager 2 (OLI-2), which collects images of Earth's landscapes in visible, near-infrared and shortwave infrared light, and the Thermal Infrared Sensor 2 (TIRS-2), which measures the temperature of land surfaces. Like its predecessors, Landsat 9 is a joint mission of NASA and the U.S. Geological Survey.

- Sentinel

The Copernicus Program is an ambitious initiative headed by the European Commission in partnership with the European Space Agency (ESA). The Sentinels are a constellation of satellites developed by ESA to operationalize the Copernicus program, which include all-weather radar images from Sentinel-1A and 1B, high-resolution optical images from Sentinel-2A and 2B, ocean and land data suitable for environmental and climate monitoring from Sentinel-3, as well as air quality data from Sentinel-5P. this project helps is study of vegetation, forest, water bodies and climate change conditions

- MODIS

The Moderate Resolution Imaging Spectroradiometer (MODIS) sensors on NASA's Terra and Aqua satellites have been acquiring images of the Earth daily since 1999, including daily imagery, 16-day BRDF-adjusted surface reflectance, and derived products such as vegetation indices and snow cover.

These sources provide both raster and vector data sources for various applications ranging from study of forest, water sources, vegetation. Marine , winds, climate, city growth and many more.

Deliverables

1. **Implementation of EO raster data Quality matrix**

References

[1] Veregin, H. (1999). Data quality parameters. *Geographical information systems*, *1*, 177-189.

[2] Caprioli, M., Scognamiglio, A., Strisciuglio, G., & Tarantino, E. (2003, August). Rules and standards for spatial data quality in GIS environments. In *Proc. 21st Int. Cartographic Conf. Durban, South Africa 10–16 August 2003*.

[3] John Byabazaire, Gregory O'Hare, Declan Delaney, Data Quality and Trust: A Perception from Shared Data in IoT In Proc. 2020 IEEE International Conference on Communications Workshops (ICC Workshops), IEEE Press, 2020.

[4] John Byabazaire, Gregory O'Hare, Declan Delaney, Using Trust as a Measure to Derive Data Quality in Data Shared IoT Deployments, 29th International Conference on Computer Communications and Networks (ICCCN), IEEE, 2020.

[5] John Byabazaire, Gregory O'Hare, Declan Delaney,, Data Quality and Trust: Review of Challenges and Opportunities for Data Sharing in IoT, Electronics 9 (12), 2083, DEc. 2020, MDPI Publishers.

[6] MGISS: https://mgiss.co.uk/

[7] GISLOUNGE :https://www.gislounge.com/

[8] NASA: //www.nasa.gov/mission_pages/landsat/overview/index.html

[9] Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003, November). Multiscale structural similarity for image quality assessment. In The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003 (Vol. 2, pp. 1398-1402). Ieee.

[10] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 586-595).

[11] Xue, W., Zhang, L., Mou, X., & Bovik, A. C. (2013). Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. IEEE transactions on image processing, 23(2), 684-695.

report

ORIGINALITY REPORT

66%	65%	33%	42%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	mgiss.co.uk Internet Source	14%
2	www.researchgate.net Internet Source	10%
3	www.opensourceagenda.com Internet Source	6%
4	www.gislounge.com Internet Source	6%
5	www.ucd.ie Internet Source	4%
6	ythi.net Internet Source	3%
7	developers.google.com Internet Source	3%
8	www.science.gov Internet Source	3%
9	idr.nitk.ac.in Internet Source	2%

10	earthobservations.org Internet Source	2%
11	Submitted to Liverpool John Moores University Student Paper	2%
12	coek.info Internet Source	2%
13	torchmetrics.readthedocs.io Internet Source	2%
14	link.springer.com Internet Source	1%
15	swmath.org Internet Source	1%
16	Submitted to University of Edinburgh Student Paper	1%
17	zvierata-keresek.com Internet Source	1%
18	Chong-Won Park. "A Practical Parity Scheme for Tolerating Triple Disk Failures in RAID Architectures", Lecture Notes in Computer Science, 2000 Publication	1%
19	eprints.um.edu.my Internet Source	1%
	www.amo.de	

20	Internet Source	1 %
21	lompocrecord.com Internet Source	<1 %
22	www.hel.fi Internet Source	<1 %
23	tel.archives-ouvertes.fr Internet Source	<1 %
24	mafiadoc.com Internet Source	<1 %
25	Submitted to Sim University Student Paper	<1 %
26	bib.irb.hr Internet Source	<1 %
27	elearning.unite.it Internet Source	<1 %
28	www.dpreview.com Internet Source	<1 %
29	"Computer Vision – ECCV 2018 Workshops", Springer Science and Business Media LLC, 2019 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		