Hindawi
Complexity
Volume 2018, Article ID 6531203, 11 pages
https://doi.org/10.1155/2018/6531203

WILEY | Hindawi

Research Article

On the Impact of Labeled Sample Selection in Semisupervised Learning for Complex Visual Recognition Tasks

Eftychios Protopapadakis,[1] Athanasios Voulodimos ,[2] and Anastasios Doulamis[1,3]

[1]*National Technical University of Athens, Zografou, 15780 Athens, Greece*
[2]*Department of Informatics and Computer Engineering, University of West Attica, Egaleo, 12243 Athens, Greece*
[3]*Institute of Communication and Computer Systems (ICCS), Zografou 15773, Athens, Greece*

Correspondence should be addressed to Athanasios Voulodimos; thanosv@mail.ntua.gr

Received 17 March 2018; Accepted 2 August 2018; Published 23 September 2018

Academic Editor: Ireneusz Czarnowski

One of the most important aspects in semisupervised learning is training set creation among a limited amount of labeled data in such a way as to maximize the representational capability and efficacy of the learning framework. In this paper, we scrutinize the effectiveness of different labeled sample selection approaches for training set creation, to be used in semisupervised learning approaches for complex visual pattern recognition problems. We propose and explore a variety of combinatory sampling approaches that are based on sparse representative instances selection (SMRS), OPTICS algorithm, k-means clustering algorithm, and random selection. These approaches are explored in the context of four semisupervised learning techniques, i.e., graph-based approaches (harmonic functions and anchor graph), low-density separation, and smoothness-based multiple regressors, and evaluated in two real-world challenging computer vision applications: image-based concrete defect recognition on tunnel surfaces and video-based activity recognition for industrial workflow monitoring.

1. Introduction

The proliferation of data generated in today's industry and economy raises the expectations for approaching towards the solutions of data-driven problems through state-of-the-art machine learning and data science techniques. One of the obstacles towards this direction, especially apparent in complex real-world applications, is the insufficient availability of ground truth, which is necessary for training and fine-tuning supervised machine learning (including deep learning) models. In this context, semisupervised learning (SSL) appears as an interesting and effective paradigm. Semisupervised learning approaches make use of both labeled and unlabeled data to create a suitable learning model given a specific problem (usually a classification problem) and related constraints. The acquisition of labeled data, for most learning problems, often requires a skilled human agent (e.g., to annotate background in an image, segment, and label video sequences for action recognition) or a physical experiment (e.g., determining the 3D structure of a protein). The cost associated with the labeling process, thus, may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive. In such situations, SSL can be of great practical value.

One major advantage is the easy implementation on existing techniques; SSL can be directly or indirectly incorporated in any machine-learning task. Semisupervised SVMs approaches are a classical example of direct usage of SSL assumptions into the minimization function [1]. Indirect utilization of SSL can be found in multiobjective optimization (MOO) frameworks [2, 3]. In MOO, we have multiple fitness evaluation functions; many of them are based on SSL assumptions. Then, from a large pool of possible solution, we peak those over the Pareto front. Thus, SSL is involved in the best individual selection procedure.

In real life, there are several fields of SSL testing, assuming that there is data availability. The work of [4] evaluates the foundation piles structural condition using graph-based approaches. A scalable graph-based approach was utilized in [5] for the initialization of a maritime surveillance system.

The SSL cluster assumption was used in [6] for the initialization of a fall detection system for elderly people. A self-training approach is adopted in [7] for industrial workflow surveillance purposes in an automobile manufacturer production line. In cultural heritage, SSL has been leveraged in [8] to develop image retrieval schemes suitable to user preferences [9].

Regarding the limitations and requirements pertaining to the selection of labeled data in SSL, there is a set of desirable properties that the utilized data should have: Firstly, representative samples are needed. The labeled samples should be able to describe (or reproduce) the original data set in the best possible way. Secondly, at least one sample per classification category is required, so that model can be able to adjust to the class properties. Finally, the existence of outliers should be considered, given that most data sets contain outliers which could lead to poor performance especially when used as labeled data (all by themselves).

In this paper, we provide a deeper insight on the effectiveness of different data sampling approaches for labeled dataset creation to be used in SSL. The data sampling approaches explored are based on sampling techniques including KenStone algorithm [10], sparse representative modeling selection (SMRS) [11], Ordering Points To Identify the Clustering Structure (OPTICS) algorithm output-based approach [12], and k-means [13] centroids and random selection. Each of the described data selection approaches is scrutinized with respect to different SSL techniques, including low-density separation [14], harmonic functions [15], pseudo-Laplacian graph regularization [16], and semisupervised regressors [17]. Our contribution lies in the investigation of two aspects on the SSL field: how can we interpret the term "few data" and how we select them in an effective manner. A preliminary version of the work presented in this paper appeared in [18]. The present work scrutinizes additional SSL techniques. Furthermore, the experimental evaluation is more thorough and extensive, including a more formal method of cluster determination, additional experiments with a different visual recognition task and dataset, and supplementary comparisons with supervised techniques as well.

The typical data selection approach in several SSL techniques, including the aforementioned ones, is, to our knowledge, the random selection of the training set. Usually, a small portion of the data, i.e., less than 40% is selected (and considered labeled); as the amount of available data increases, the fraction of the required labeled instances decreases [19, 20]. At this point, two problems become apparent: (i) the number of selected instances is subjective to the expert's view and (ii) random selection does not guarantee that the major sources of variance appear in the labeled data set. In this paper, we adopt data-driven approaches for data sampling, trying to identify appropriate sampling selection techniques for SSL models.

The remainder of this paper is structured as follows: In Section 2, we first briefly present four known techniques used in the bibliography for clustering and/or sampling, which we then combine to derive seven data selection approaches. The efficacy of these approaches as labeled data generators for the SSL techniques presented in Section 3 will be evaluated in the context of two complex multiclass visual classification problems, i.e., defect recognition on concrete tunnel surfaces and activity recognition in industrial workflow monitoring. The related experimental results are presented and discussed in Section 4. Finally, Section 5 concludes the paper with a summary of findings.

2. Labeled Sample Selection Approaches for Training Data Set Creation

Given a set of feature values for a data sample, a two-step process is adopted in the analysis conducted in this study. The first step involves data sampling, i.e., the selection of the most descriptive representatives in the available data set. The second step employs popular data mining algorithms; i.e., predictive models are trained over the descriptive subsets of the previous step.

The main purpose of data sampling is the selection of appropriate representative samples to provide a good training set and, thus, improve the classification performance of predictive models. In this section, we present seven (7) data sampling approaches, which are based on the combination or adaptation of four (4) main known sampling techniques [21].

2.1. Main Techniques. The most important factor in data selection is the definition of distance function. For any two given data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, $\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{m}}$ let $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ denote the distance between them. Let $\boldsymbol{A} \in \mathbb{R}^{m \times m}$ be a symmetric matrix. The distance measure defined as

$$d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{A} (\boldsymbol{x}_i - \boldsymbol{x}_j)}. \quad (1)$$

Most of the proposed approaches are based on the Euclidean distance (i.e., $\boldsymbol{A} = \boldsymbol{I}$). Sampling algorithms are used over the entire data set $\boldsymbol{\mathcal{X}}$ and create a new set, $\boldsymbol{\mathcal{X}}_r \subset \boldsymbol{\mathcal{X}}$, according to the data relationships, as described by the distance among them. In this study, we need at least one observation from every possible class.

2.1.1. OPTICS Algorithm. Ordering Points to Identify the Clustering Structure (OPTICS) is an algorithm for finding density-based clusters in spatial data [22], i.e., detect meaningful clusters in data of varying density. The points of the database are (linearly) ordered such that points which are spatially closest become neighbors in the ordering.

OPTICS requires two parameters: the maximum distance (radius) to consider (ε) and the number of points required to form a cluster (*MinPts*) MinPts. A point p is a core point if at least MinPts points are found within its ε-neighborhood, $N_\varepsilon(p)$. Once the initial clustering is formed, we may proceed with any sampling approach (e.g., random selection among clusters).

2.1.2. k-Means Algorithm. k-means clustering [13] aims to partition n observations into k clusters, such that each observation is assigned to the cluster it is most similar to (with the cluster centroid serving as a prototype of the cluster). It is a classical approach that can be implemented in many ways

and for various distance metrics. The main drawback is that the number of clusters should be known a priori.

2.1.3. Sparse Modeling for Representative Selection. Sparse modeling representative selection (SMRS) focuses on the identification of representative objects through the solution of the following optimization problem [11]:

$$\begin{aligned} \min \quad & \lambda \|\boldsymbol{C}\|_{1,q} + \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{C}\|_F^2 \\ \text{s.t.} \quad & \mathbf{1}^T\boldsymbol{C} = \mathbf{1}^T, \end{aligned} \tag{2}$$

where $\boldsymbol{X}$ and $\boldsymbol{C}$ refer to data points and coefficient matrix, respectively. This optimization problem can also be viewed as a compression scheme, where we want to choose a few representatives that can reconstruct the available data set.

2.1.4. Kennard–Stone Algorithm. Using the classic KenStone algorithm, we can cover the experimental area in a uniform way, since it provides a flat data distribution. The algorithm's main idea is that to select the next sample, it opts for the sample whose distance to those that have been previously chosen (called calibration samples) is the greatest.

Therefore, among all possible points, the algorithm selects the point which is furthest from those already selected and adds it to the set of calibration points. To this end, the distance is calculated between each candidate point $\boldsymbol{x}_0$ to each point $\boldsymbol{x}$ which has already been selected. In the sequel, we determine which one is the smallest, i.e., $\min_i d(\boldsymbol{x}, \boldsymbol{x}_0)$. Among these, we choose the point for which the distance is maximal:

$$d_{\text{selected}} = \max_{i_0} \left(\min_i d(\boldsymbol{x}_i, \boldsymbol{x}_0) \right). \tag{3}$$

2.2. Combinatory Sampling Approaches. The primary goal of sampling approaches is the removal of redundant and uninformative data. Using the algorithms described earlier in Section 2.1 as a basis, we propose six (6) combinatory sampling approaches. A brief description of each one, along with the baseline random selection method, follows:

(i) *OPTICS extrema*: after employing the OPTICS algorithm on the entire data set, the calculated reachability distances are plotted in the same order as data were processed. Over the generated waveform, we locate local maxima and minima. All the identified extrema cases are considered as labeled instances and the rest as unlabeled. This approach results in a very limited training set.

(ii) *Sparse modeling representative selection (SMRS)*: the SMRS technique is employed over the entire data set, resulting in a very limited training set, although larger than the one obtained with OPTICS. In contrast to OPTICS, the selected points are located only on the exterior cell of the available data volume.

(iii) *Combination of k-means and SMRS (k-means SMRS)*: we first divide the set into k subclusters. For each subcluster, we run the SMRS algorithm to get the representative samples among each subcluster. As such, the outcome provides points surrounding each subcluster. The number of clusters, k, was defined using the Silhouette score for all k values, $k \in [2, \mathcal{u} + 4]$, where $\mathcal{u}$ is a heuristic approach estimating the number of clusters, defined as $\mathcal{u} = \lceil \sqrt{n/2} \rceil$, and n denotes the number of available data instances (observations).

(iv) *Combination of OPTICS and SMRS (OPTICS-SMRS)*: SMRS is performed to the subclusters obtained through the OPTICS algorithm. This approach is similar to the work of [19]. A subset is created of representative samples from each subcluster obtained by OPTICS algorithm. The minimum number of data within a cluster, required by OPTICS, was defined as MinPts = min $(\lfloor n/k \rfloor, 8)$.

(v) *Kennard and Stone (KenStone) sampling data points*: after executing the KenStone algorithm, we have data entries spanning uniformly the entire data space.

(vi) *Random selection*: a random selection that picks $p\%$ of the available data as training data, this is the baseline data selection method used in the context of most SSL techniques.

(vii) *Improved random selection*: an alternative approach is the creation of k clusters (using k-means) and a random selection of n_k samples from each cluster (k-means random). It is an improvement of random selection, without involving any advanced techniques. Similar instances are likely to be clustered together. Thus, the few random samples from each cluster are expected to provide adequate information over the data set.

All of the proposed approaches are applied over all available data, labeled or not. As such, it is possible for many of the selected training data to be unlabeled. In that case, an expert would be summoned to annotate the selected data, as would have been the case in any annotation attempt. However, in this case, the annotation effort will be less considerable compared to traditional supervised approaches, which use a significantly higher percentage of the available data for training purposes.

3. Semisupervised Learning Techniques

In this work, four of the most popular types of SSL techniques will be considered: two graph-based approaches, along with low-density separation, and multiple smoothness assumption-related regressors.

3.1. Graph-Based Approaches. Graph-based semisupervised methods define a graph over the entire data set, $\boldsymbol{X} = \boldsymbol{X}_{\text{L}} \cup \boldsymbol{X}_{\text{U}}$, where $\boldsymbol{X}_{\text{L}} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_l, \boldsymbol{y}_l)\}$ is the labeled data set

and $\boldsymbol{X}_{\mathrm{U}} = \{\boldsymbol{x}_{l+1}, \ldots, \boldsymbol{x}_{l+u}\}$ the unlabeled data set. Feature vectors, $\boldsymbol{x}_i \in \mathbb{R}^m$, $i = 1, \ldots, l+u$, are available for all the observations and $\boldsymbol{y}_i \in \mathbb{R}^C$, $i = 1, \ldots, l$ are the corresponding classes of the labeled ones, in a vector form; C denotes the available classes.

The nodes represent the labeled and unlabeled examples in the dataset; edges reflect the similarity among examples. In order to quantify the edges (i.e., assign a similarity value), an adjacency matrix $\boldsymbol{\mathcal{A}}$ is calculated, where

$$\mathcal{A}_{ij} = \begin{cases} 1 & \text{if } \boldsymbol{x}_i \text{ close to } \boldsymbol{x}_j \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

Practically, each label is only connected to its k closest labels, so that $\sum_{j=0}^{n} \mathcal{A}_{ij} = k$. The information of the labeled nodes propagates to the unlabeled nodes via paths defined on existing edges provided by $\boldsymbol{\mathcal{A}}$.

Graph methods are nonparametric, discriminative, and transductive in nature. Intuitively speaking, in a graph that various data points are connected, the greater the similarity, the greater the probability of having similar labels. Thus, the information (of labels) propagates from the labeled points to the unlabeled ones. These methods usually assume label smoothness over the graph. That is, if two instances are connected by a strong edge, their labels tend to be the same.

3.1.1. Harmonic Functions. An indicative paradigm of graph-based SSL is the harmonic function approach [23]. This approach estimates a function f on the graph which satisfies two conditions. Firstly, f has the same values as given labels on the labeled data, i.e., $f(\boldsymbol{x}_i) = \boldsymbol{y}_i$, $i = 1, \ldots, l$. Secondly, f satisfies the weighted average property on the unlabeled data:

$$f(\boldsymbol{x}_j) = \frac{\sum_{k=1}^{l+u} w_{jk} f(\boldsymbol{x}_j)}{\sum_{k=1}^{l+u} w_{jk} f}, \quad j = l+1, \ldots, l+u, \tag{5}$$

where w_{ij} denotes the edge weight. Those two conditions lead to the following problem:

$$\begin{aligned} \min_{f: f(\boldsymbol{x}) \in \mathbb{R}} \quad & \sum_{i,j=1}^{l+u} w_{ij} \left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)\right)^2 \\ \text{s.t.} \quad & f(\boldsymbol{x}_i) = \boldsymbol{y}_i, \; i = 1, \ldots, l. \end{aligned} \tag{6}$$

The problem has an explicit solution, which allows a soft label estimation for all the edges of the graph, i.e., investigated cases.

3.1.2. Anchor Graph. Anchor graph estimates a labeling prediction function $f : \mathbb{R}^m \to \mathbb{R}$ defined on the samples of $\boldsymbol{X}$; by using a subset $\boldsymbol{\mathcal{U}} = \{\boldsymbol{u}_k\}_k^p \subset \boldsymbol{X}_L$ of the labeled data, the label prediction function can be expressed as a convex combination [16]:

$$f(\boldsymbol{x}_i) = \sum_{k=1}^{p} Z_{ik} \cdot g(\boldsymbol{u}_k), \tag{7}$$

where Z_{ik} denotes sample-adaptive weights, which must satisfy the constraints $\sum_{k=1}^{l} Z_{ik} = 1$ and $Z_{ik} \geq 0$ (convex combination constraints). By defining vectors $\boldsymbol{g}$ and $\boldsymbol{a}$, respectively, as $\boldsymbol{g} = [g(\boldsymbol{f}_1), \ldots, g(\boldsymbol{f}_n)]^{\boldsymbol{T}}$ and $\boldsymbol{a} = [g(\boldsymbol{x}_1), \ldots, g(\boldsymbol{x}_p)]^{\boldsymbol{T}}$, (7) can be rewritten as $\boldsymbol{g} = \boldsymbol{Z}\boldsymbol{\alpha}$ where $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$.

The designing of matrix $\boldsymbol{Z}$, which measures the underlying relationship between the samples of $\boldsymbol{X}_U$ and samples $\boldsymbol{X}_L$, is based on weight optimization; i.e., nonparametric regression. Thus, the reconstruction for any data point is a convex combination of its closest representative samples.

Nevertheless, the creation of matrix $\boldsymbol{Z}$ is not sufficient, as it does not assure a smooth function $\boldsymbol{g}$. There is always the possibility of inconsistencies in segmentation, i.e., different samples with almost identical attributes belong to different classes. In order to deal with such cases, the following SSL framework is employed:

$$\min_{\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_c]} \quad \mathcal{Q}(\boldsymbol{A}) = \frac{1}{2} \|\boldsymbol{Z}\boldsymbol{A} - \boldsymbol{Y}\|_F^2 + \frac{\gamma}{2} \operatorname{trace}\left(\boldsymbol{A}^{\boldsymbol{T}} \widehat{\boldsymbol{L}} \boldsymbol{A}\right), \tag{8}$$

where $\widehat{\boldsymbol{L}} = \boldsymbol{Z}^T \boldsymbol{L} \boldsymbol{Z}$ is a memory-wise and computationally tractable alternative of the Laplacian matrix $\boldsymbol{L}$. Matrix $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_c] \in \mathbb{R}^{p \times c}$ is the soft label matrix for the representative samples, in which each column vector accounts for a class. The matrix $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_c] \in \mathbb{R}^{n \times c}$ is a class indicator matrix on ambiguously labeled samples with $Y_{ij} = 1$ if the label l_i of sample i is equal to j and $Y_{ij} = 0$ otherwise.

The Laplacian matrix $\boldsymbol{L}$ is calculated as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, where $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ is a diagonal degree matrix and $\boldsymbol{W}$ is approximated as $\boldsymbol{W} = \boldsymbol{Z}\boldsymbol{\Lambda}^{-1}\boldsymbol{Z}^T$. Matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{p \times p}$ is defined as $\boldsymbol{\Lambda} = \sum_{i=1}^{n} Z_{ik}$. The solution of (8) has the form

$$\boldsymbol{A}^* = \left(\boldsymbol{Z}^T \boldsymbol{Z} + \gamma \widehat{\boldsymbol{L}}\right) \square^T \boldsymbol{Y}. \tag{9}$$

Each sample label is, then, given by

$$\widehat{l}_i = \arg \max_{j \in \{1, \ldots, c\}} \frac{\boldsymbol{Z}_i a_j}{\lambda_j}, \tag{10}$$

where $\boldsymbol{Z}_i \in \mathbb{R}^{1 \times p}$ denotes the i-th row of $\boldsymbol{Z}$, and the normalization factor $\lambda_j = \boldsymbol{1}^T \boldsymbol{Z}\, \boldsymbol{\alpha}_j$ balances skewed class distributions.

3.2. Low-Density Separation. The low-density separation assumption pushes the decision boundary in regions where there are few data points (labeled or unlabeled). The most common approach to achieving this goal is to use a maximum margin algorithm such as support vector machines. The method of maximizing the margin for unlabeled as well as labeled points is called the transductive SVM (TSVM). However, the corresponding problem is nonconvex and thus difficult to solve [24].

Low-density separation (LDS) is a combination of TSVMs [25], trained using gradient descend, and traditional SVMs using an appropriate kernel defined over a graph using SSL assumptions [14]. Like the SVM approach, the TSVM maximizes the class-separating margin.

The problem can be stated in the following form, which allows for a standard gradient-based approach:

$$\min_{\boldsymbol{w},b}\left[\frac{1}{2}\boldsymbol{w}^2 + C\sum_{i=1}^{l} L^2\left(y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right)\right) + C^*\sum_{j=l+1}^{l+u} L^*\left(\left|\boldsymbol{w}^T\boldsymbol{x}_j + b\right|\right)\right], \tag{11}$$

where $\boldsymbol{w} \in \mathbb{R}^n$ is the parameter vector that specifies the orientation and scale of the decision boundary and $b \in \mathbb{R}$ is an offset parameter. The above formulation exploits both labeled $\boldsymbol{X}_{\mathrm{L}}$ and unlabeled $\boldsymbol{X}_{\mathrm{U}}$ data. Finally, let us denote as $L(t) = \max(0, 1-t)$ and $L^*(t) = \exp(-3t^2)$.

Such a formulation allows the use of a nonlinear kernel, calculated over a fully connected matrix, $\boldsymbol{W}$, which is formed as $w_{ij} = \exp(\rho - dist(i,j)) - 1$. Dijkstra's algorithm is employed to compute the shortest path lengths, $d_{\mathrm{SP}}(i,j)$ for all pairs of points. The matrix $\mathscr{D}$ of squared ρ-path distances is calculated for all pairs of points as

$$\mathscr{D}_{ij} = \left(\frac{1}{\rho}\log\left(1 + d_{\mathrm{SP}}(i,j)\right)\right)^2. \tag{12}$$

The final step towards the kernel's creation involves multidimensional scaling [23], or MDS, to find a Euclidean embedding of $\mathscr{D}^{\rho}$ (in order to obtain a positive definite kernel). The embedding found by the classical MDS are the eigenvectors corresponding to the positive eigenvalues $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T = -\boldsymbol{H}\mathscr{D}^{\rho}\boldsymbol{H}$, where $H_{ij} = \delta_{ij} - 1/(l+u)$. The final representation of $\boldsymbol{x}_i$ is $x_{ik} = U_{ik}\sqrt{\lambda_k}$, $1 \le k \le p$.

3.3. Semisupervised Regression. The safe semisupervised regression (SAFER) approach [17] tries to learn a prediction from several semisupervised regressors. Specifically, let $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_b\}$ be multiple SSR predictions and $\boldsymbol{f}_0$ be the prediction of a direct supervised learner, where $\boldsymbol{f}_i \in \mathbb{R}^U$, $i = 1, \ldots, r$ and r refers to the number of regressors. Supposing there is no knowledge with regard to the reliabilities of learners, SAFER optimizes the performance gain of $g(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_b, \boldsymbol{f}_0)$ against $\boldsymbol{f}_0$, when the weights of SSR learners come from a convex set.

The problem lies in the solution of the following equation:

$$\max_{\boldsymbol{f}\in\mathbb{R}^U}\min_{\substack{\boldsymbol{\alpha}\in\\ \mathscr{M}}}\quad \sum_{i=1}^{r}\alpha_i\left(\|\boldsymbol{f}_0 - \boldsymbol{f}_i\|^2 - \|\boldsymbol{f} - \boldsymbol{f}_i\|^2\right), \tag{13}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_r]$, $\alpha_i \ge 0$, are the weights of individual regressors. Equation (12) is concave to $\boldsymbol{f}$ and convex to $\boldsymbol{\alpha}$. Thus, it is recognized as saddle-point convex-concave optimization [26].

4. Experimental Evaluation

We will hereby examine the applicability and effectiveness of each of the above-described data selection techniques for the SSL approaches presented. SSL is particularly useful in cases where there is limited availability of labeled data and/or the creation of appropriately sized labeled data sets requires a prohibitive amount of resources, as is the case in real-world visual classification problems. Two prominent examples of such applications are (a) automated image-based detection and classification of defects on concrete surfaces in the context of visual inspection of tunnels [27] and (b) human activity recognition from video, e.g., the monitoring of workflow in industrial assembly lines [28, 29].

MATLAB software has been used for the implementation of the proposed approaches. The SSL approaches code, i.e., Harmonic functions, Anchor graph, LDS, and SAFER, were provided by the corresponding authors of [14, 16, 17, 23]. OPTICS, KenStone, and SMRS as well as code implementations were provided by [11, 22, 30], respectively.

4.1. Defect Recognition on Tunnel Concrete Surfaces. The tunnel defect recognition dataset (henceforth referred to in this paper as the *Tunnel dataset*) consists of images acquired by a robot inside a tunnel of Egnatia Motorway, in Greece, in the context of ROBO-SPECT project [27]. Images were used for detecting and recognizing defects on the concrete surfaces. Raw captured tunnel and annotated ground truth images of resolution 600×900 pixels were provided. Figure 1 shows some examples from the Tunnel dataset displaying cracked areas on the concrete surface.

To represent each pixel, we use the same low-level feature extraction techniques as in [27]; in particular, each pixel p_{xy} is described by a feature vector $\boldsymbol{s}_{xy} = [s_{1,xy}, \ldots, s_{k,xy}]^T$, where s are scalars corresponding to the presence and magnitude of the low-level features detected at location (x, y). Figure 2 displays the extracted low-level features. Feature vectors along with the class labels of every pixel are used to form a data set. There are five different classes of defects: (1) crack, (2) staining, (3) spalling, (4) calcium leaching, and (5) unclassified.

We, hereby, briefly describe the features used to form vector $\boldsymbol{s}_{xy}$. First, we take the edges denoted by a pixel-wise multiplication of the Canny and Sobel operators. Secondly, frequency is calculated as $\mathscr{F}_I = \nabla^2 I$. Thirdly, we calculate the entropy in order to separate homogenous regions from textured ones. Texture was described using twelve Gabor filters with orientations 0°, 30°, 60°, and 90° and frequencies 0.0, 0.1, and 0.4. The Histogram of Oriented Gradients (HOG) was also calculated. By combining these features with the raw pixels' intensity, feature vector $\boldsymbol{s}_{xy}$ takes the form of a 1×17 vector containing visual information that characterizes each one of the image pixels.

A typical K-fold validation approach is adopted, resulting in eight (approximately) equal partitions, i.e., disjoint subsets, of the n observations. The training set size is limited at 3% of sample population, when random techniques and KenStone algorithm were applied.

4.2. Activity Recognition from Video for Industrial Workflow Recognition. Action or activity recognition from video is a very popular computer vision application. A significant application domain is automatic video surveillance, e.g., for safety, security, and quality assurance reasons. In this

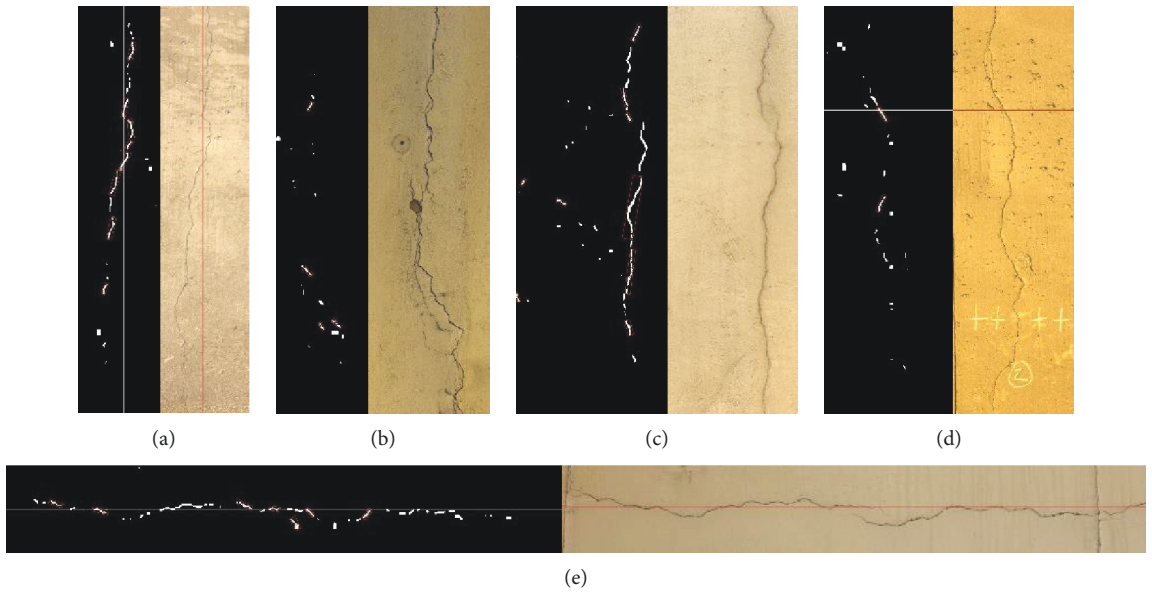

Figure 1: Examples of cracked areas from the Tunnel dataset.

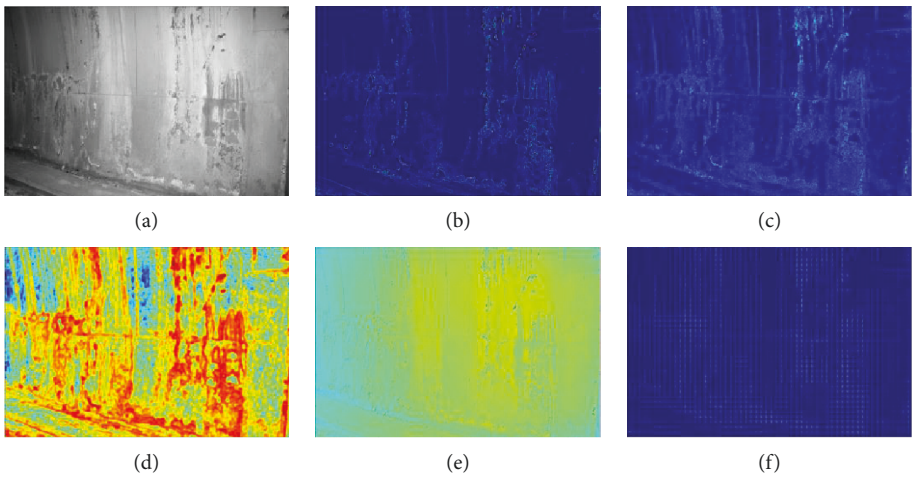

Figure 2: Illustration of the extracted low-level features in the Tunnel dataset: (a) original image, (b) edges, (c) frequency, (d) entropy, (e) texture, and (f) HOG.

experiment, we will make use of real-world video sequences from the surveillance camera of a major automobile manufacturer (NISSAN) [31], captured in the context of the SCOVIS EU project in the publicly available Workflow Recognition (WR) dataset [32].

The production cycle on the industrial line included tasks of picking several parts from racks and placing them on a designated cell some meters away, where welding took place. Each of the above tasks was regarded as a class of behavioral patterns that had to be recognized. The activities (tasks) we were aiming to model in the examined application are briefly the following:

(1) One worker picks part #1 from rack #1 and places it on the welding cell

(2) Two workers pick part #2a from rack #2 and place it on the welding cell

(3) Two workers pick part #2b from rack #3 and place it on the welding cell

(4) One worker picks up parts #3a and #3b from rack #4 and places them on the welding cell

(5) One worker picks up part #4 from rack #1 and places it on the welding cell

(6) Two workers pick up part #5 from rack #5 and place it on the welding cell

(7) Workers were idle or absent (null task)

The WR dataset includes twenty full cycles, each containing occurrences of the above tasks. Figure 3 depicts a typical example of an execution of Task 2. The visual classification problem in this case is to automatically recognize which task is executed at every time instance.

Figure 3: Indicative example of key-frames corresponding to the execution of a task (Task 2).

Table 1: Illustration of the training set data size per sampling approach (averages over all tests).

Row labels	KenStone	kmeansRandom	kmeansSMRS	OPTICS extrema	OPTICS-SMRS	Random	SMRS	Entire set
WR	156	181.25	422.37	289.75	532.39	156	23.62	5199
Tunnel	36.37	38	37.75	55	141.76	36.37	14.12	1200

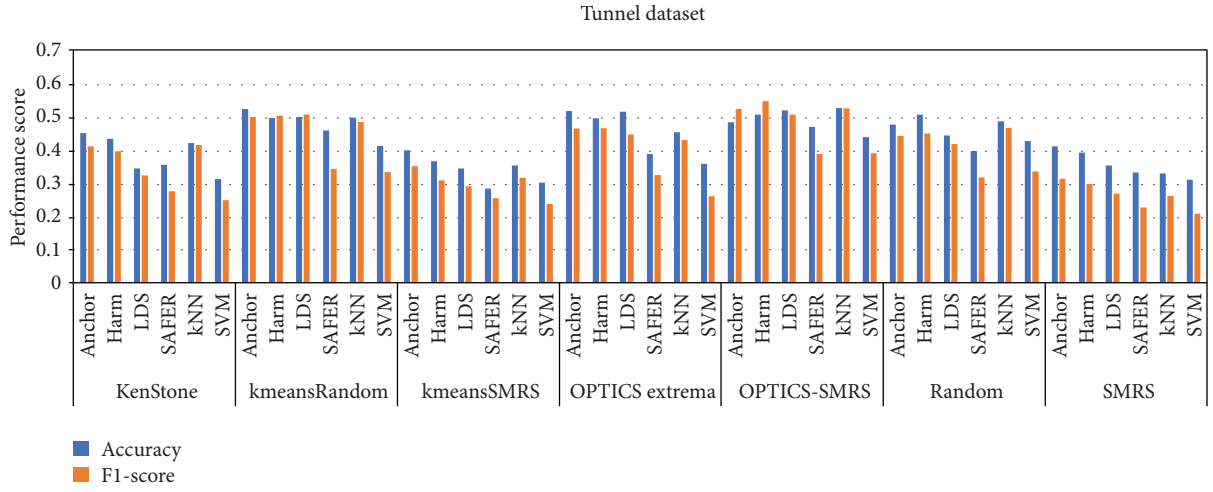

Figure 4: Performance scores for all data selection and SSL combinations (Tunnel dataset).

In all video segments, holistic features such as Pixel Change History (PCH) are used. These features remedy the drawbacks of local features, while also necessitating a far less tedious computational procedure for their extraction [33]. A very positive attribute of such representations is that they can easily capture the history of a task that is being executed. These images can then transform to a vector-based representation using the Zernike moments (up to sixth order, in our case) as applied in [33, 34]. The video features, once exported, had a two-dimensional matrix representation of the form $m \times l$, where m denotes the size of the $1 \times m$ vectors created using Zernike moments and l the number of such vectors.

4.3. Experimental Results. Each of the seven data sampling approaches described in Section 2.2 was paired with each of the four SSL techniques presented in Section 3 as well as two well-known supervised approaches, i.e., SVM and kNN, resulting in 42 combinations in total. Table 1 illustrates the training data set size generated in the case of each data selection approach applied for the two datasets. It is interesting to note here that the OPTICS-SMRS approach provides significantly more data than any other approach.

The classification results in terms of averaged accuracy and F-measure for each combination are depicted in Figure 4 for defect recognition (Tunnel dataset) and Figure 5 for activity recognition (WR dataset). At first look, it appears that among SSL techniques, it is harmonic functions that tend to provide higher accuracy rates, while concerning data sampling approaches, cluster-based selection

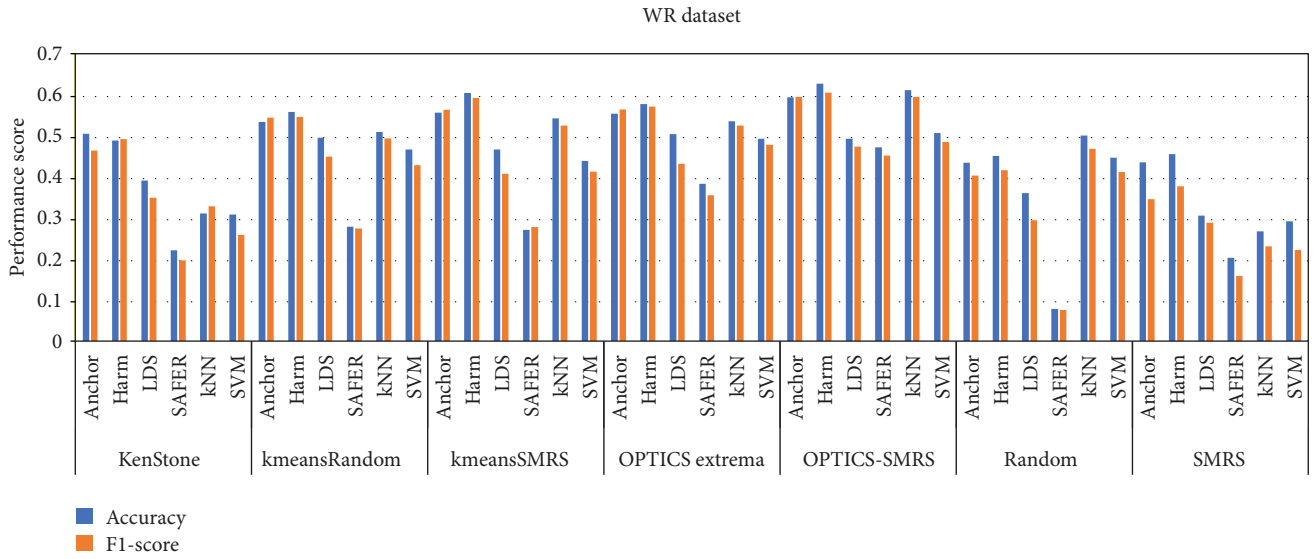

Figure 5: Performance scores for all data selection and SSL combinations (WR dataset).

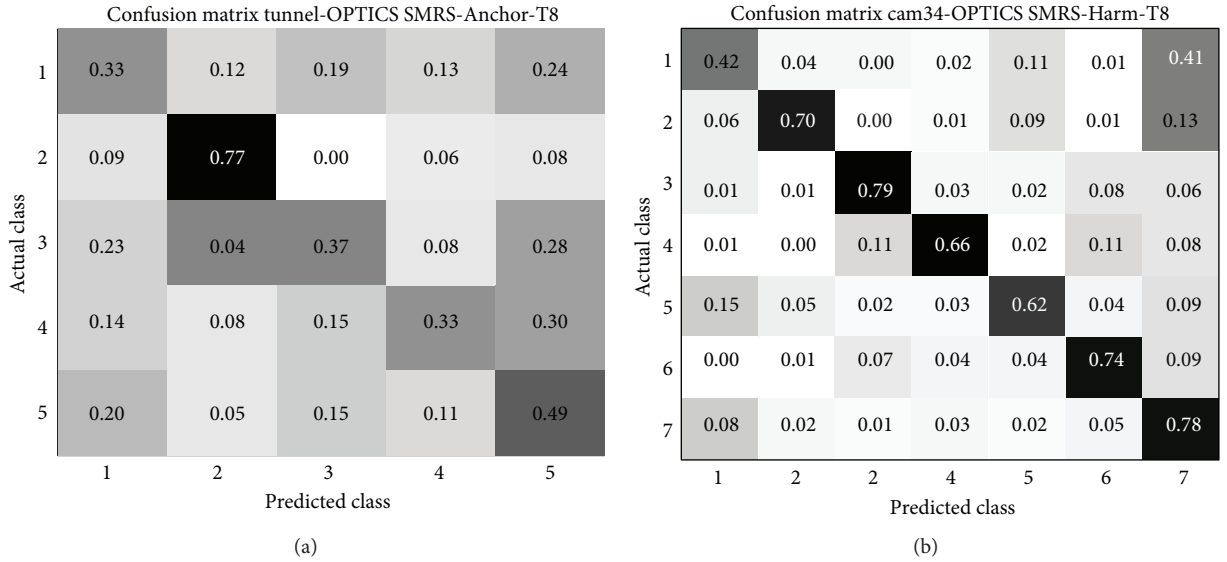

Figure 6: Confusion matrices for OPTICS-SMRS sampling in (a) Tunnel dataset, using anchor graph and (b) WR dataset, using harmonic functions.

(centroid or density-based) appears to give overall better results. Figure 6 provides an example confusion matrix for each visual recognition problem, acquired for OPTICS-SMRS data selection method.

Figure 4 illustrates the performance of the combinatory models in the tunnel surface defect recognition task. Cluster-based selection (OPTICS-SMRS followed by k-means random) appears to be the data selection techniques that lead to the best performance rates. Additionally, graph-based classifiers tend to perform better in most cases. The low performance scores for all the cases can be put down to the extremely challenging nature of the problem, as well as the feature quality; it is very likely for various defect types to have similar feature values when using low-level features [35].

Figure 5 illustrates the performance for the combinatory models in the WR dataset. Again, OPTICS-SMRS sampler appears to lead to the best performance rates, especially when using harmonic functions as SSL technique. It is interesting to note that, when using most of the proposed data selection techniques for training set creation, graph-based SSL techniques (harmonic functions and anchor graph) outperform not only the remaining SSL techniques but also the supervised methods examined, i.e., kNN and SVM. This can be explained by the lower number of training samples used compared to the usual training set sizes in such supervised learning methods.

4.4. Statistical Tests. In order to derive further conclusions regarding the results and the relative performance of the

Table 2: ANOVA results.

Source	Sum sq.	d.f.	Mean sq.	F	p value
Sampling	3.5488	6	0.5915	167.0981	0
Classifier	3.1569	5	0.6314	178.2768	0
Number of classes	0.2687	1	0.2687	75.9157	0
Sampling × classifier	0.3766	30	0.0126	3.5469	0
Sampling × num of classes	0.7855	6	0.1309	36.9865	0
Classifier × num of classes	0.4715	5	0.0943	26.6411	0
Error	2.1769	615	0.0035		
Total	10.7920	668			

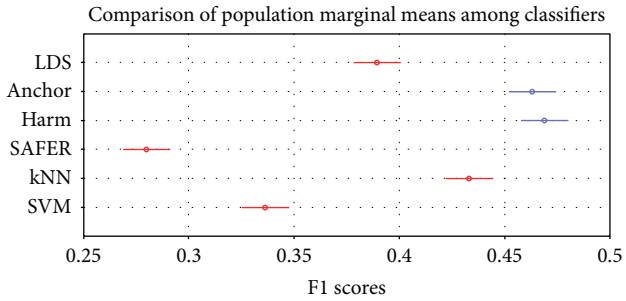

Figure 7: F1 scores by classification method.

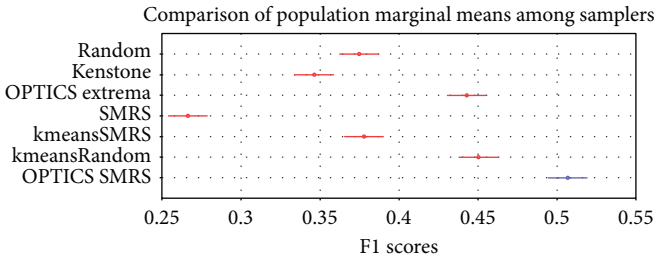

Figure 8: F1 scores by data selection method.

technique combinations explored, we performed an analysis of variance (ANOVA) on the F1 scores for the test samples. ANOVA permits the statistical evaluation of the effects of the two main design factors of this analysis (i.e., the sampling schemes and the SSL techniques). As shown in Table 2, both the sampling scheme and the choice of classifier are strongly significant for explaining variations in F1 scores. The dataset impact is also significant; i.e., performance variations should be expected in other datasets.

Apart from the above basic ANOVA results, we use the Tukey honest significant difference (HSD) post hoc test so as to derive conclusions about the best performing approaches, taking into account the statistical significance of the variations in the values of metrics presented. Figures 7 and 8 illustrate the results for the SSL techniques and the sampling schemes, respectively, for the entirety of experiments conducted.

As far as SSL techniques are concerned, harmonic functions and anchor graph appear to have a statistically significant superiority over all alternatives. The outcome verifies previous analysis outcomes (see Figures 4 and 5) suggesting that graph-based approaches result in better rates compared to the other SSL (or even supervised learning) alternatives (see Figure 7). The low overall performance scores in the comparison of learning techniques can be explained by the challenging nature of both examined problems as well as by the fact that all configurations have been taken into consideration including those yielding very low performance rates.

Finally, as regards data selection techniques, we observe that the OPTICS-based approach combined with SMRS creates training sets that lead to clearly the highest performance rates among all examined techniques, including the traditionally used random sampling. Furthermore, we can see that cluster-based samplers in general yield results that are at least as good as random sampling. On the other hand, SMRS alone provides results significantly worse than all competing schemes.

5. Conclusion

The creation of a training set of labeled data is of great importance for semisupervised learning methods. In this work, we explored the effectiveness of different data sampling approaches for labeled data generation to be used in SSL models in the context of complex real-world computer vision applications. We compared seven sampling approaches, some of which we proposed in this paper, all based on OPTICS, k-means, SMRS, and KenStone algorithm. The proposed data selection approaches were used to create labeled data sets to be used in the context of four SSL techniques, i.e., anchor graph, harmonic functions, low-density separation, and semisupervised regression. Extensive experiments were carried out in two different and very challenging real-world visual recognition scenarios: image-based concrete defect recognition on tunnel surfaces and video-based activity recognition for industrial workflow monitoring. The results indicate that SSL data selection schemes, using density-based clustering prior to sampling, such as a combination of OPTICS and SMRS algorithms, provide better performance results compared to traditional sampling approaches, such as random selection. Finally, as regards the SSL techniques studied, graph-based approaches (harmonic functions and anchor graph) appeared to have a statistically significant superiority for the two visual recognition problems examined.

Data Availability

The WR dataset is publicly available as described in [30]. The Tunnel dataset was created for the research activities of the ROBO-SPECT EU project (http://www.robo-spect.eu) and is not publicly available due to confidentiality restrictions. However, a small number of partially annotated images can be provided by the authors upon request.

Disclosure

Part of the work presented in this paper has been included in the doctoral thesis of Dr. Eftychios Protopapadakis titled "Decision Making via Semi-Supervised Machine Learning Techniques."

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

The research leading to these results has received funding from the European Commission's H2020 Research and Innovation Programme under Grant Agreement no. 740610 (STOP-IT project).

References

[1] W.-J. Chen, Y.-H. Shao, and N. Hong, "Laplacian smooth twin support vector machine for semi-supervised classification," *International Journal of Machine Learning and Cybernetics*, vol. 5, no. 3, pp. 459–468, 2014.

[2] E. Protopapadakis, A. Voulodimos, and N. Doulamis, "An investigation on multi-objective optimization of feedforward neural network topology," in *2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pp. 1–6, Larnaca, Cyprus, 2017.

[3] A. K. Alok, S. Saha, and A. Ekbal, "Semi-supervised clustering for gene-expression data in multiobjective optimization framework," *International Journal of Machine Learning and Cybernetics*, vol. 8, no. 2, pp. 421–439, 2017.

[4] E. Protopapadakis, M. Schauer, E. Pierri et al., "A genetically optimized neural classifier applied to numerical pile integrity tests considering concrete piles," *Computers & Structures*, vol. 162, pp. 68–79, 2016.

[5] K. Makantasis, E. Protopapadakis, A. Doulamis, and N. Matsatsinis, "Semi-supervised vision-based maritime surveillance system using fused visual attention maps," *Multimedia Tools and Applications*, vol. 75, no. 22, pp. 15051–15078, 2016.

[6] K. Makantasis, E. Protopapadakis, A. Doulamis, N. Doulamis, and N. Matsatsinis, "3D measures exploitation for a monocular semi-supervised fall detection system," *Multimedia Tools and Applications*, vol. 75, no. 22, pp. 15017–15049, 2016.

[7] E. Protopapadakis, A. Doulamis, K. Makantasis, and A. Voulodimos, *A Semi-Supervised Approach for Industrial Workflow Recognition*, INFOCOMP, 2012.

[8] E. Protopapadakis and A. Doulamis, "Semi-supervised image meta-filtering using relevance feedback in cultural heritage applications," *International Journal of Heritage in the Digital Era*, vol. 3, no. 4, pp. 613–627, 2014.

[9] A. S. Voulodimos and C. Z. Patrikakis, "Quantifying privacy in terms of entropy for context aware services," *Identity in the Information Society*, vol. 2, no. 2, pp. 155–169, 2009.

[10] R. W. Kennard and L. A. Stone, "Computer aided design of experiments," *Technometrics*, vol. 11, no. 1, pp. 137–148, 1969.

[11] R. Vidal, G. Sapiro, and E. Elhamifar, "See all by looking at a few: sparse modeling for finding representative objects," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1600–1607, Providence, RI, USA, 2012.

[12] E. Protopapadakis, A. Voulodimos, A. Doulamis, N. Doulamis, D. Dres, and M. Bimpas, "Stacked autoencoders for outlier detection in over-the-horizon radar signals," *Computational Intelligence and Neuroscience*, vol. 2017, Article ID 5891417, 11 pages, 2017.

[13] J. Wu, "Cluster analysis and K-means clustering: an introduction," in *Advances in K-means Clustering*, pp. 1–16, Springer, 2012.

[14] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," *AISTATS*, vol. 2005, pp. 57–64, 2005.

[15] X. Zhu, J. Lafferty, and Z. Ghahramani, "Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions," in *ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining*, vol. 3, pp. 58–65, AAAI Press, 2003.

[16] W. Liu, J. He, and S.-F. Chang, "Large graph construction for scalable semi-supervised learning," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 679–686, Haifa, Israel, 2010.

[17] Y.-F. Li, H.-W. Zha, and Z.-H. Zhou, "Learning safe prediction for semi-supervised regression," *AAAI*, vol. 2017, pp. 2217–2223, 2017.

[18] E. Protopapadakis, A. Voulodimos, and A. Doulamis, "Data sampling for semi-supervised learning in vision-based concrete defect recognition," in *2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pp. 1–6, Larnaca, Cyprus, 2017.

[19] I. Triguero, S. García, and F. Herrera, "Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study," *Knowledge and Information Systems*, vol. 42, no. 2, pp. 245–284, 2015.

[20] N. F. F. Da Silva, L. F. S. Coletta, and E. R. Hruschka, "A survey and comparative study of tweet sentiment analysis via semi-supervised learning," *ACM Computing Surveys*, vol. 49, no. 1, pp. 1–26, 2016.

[21] E. Protopapadakis, "Decision making via semi-supervised machine learning techniques," 2016, http://arxiv.org/abs/1606.09022.

[22] M. Daszykowski, B. Walczak, and D. L. Massart, "Looking for natural patterns in analytical data. 2. Tracing local density with OPTICS," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 3, pp. 500–507, 2002.

[23] X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pp. 912–919, Washington, DC, USA, 2003.

[24] A. Singla, S. Patra, and L. Bruzzone, "A novel classification technique based on progressive transductive SVM learning," *Pattern Recognition Letters*, vol. 42, pp. 101–106, 2014.

[25] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive SVM for semisupervised classification of remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 11, pp. 3363–3373, 2006.

[26] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, vol. 87, Springer Science & Business Media, 2013.

[27] K. Makantasis, E. Protopapadakis, A. Doulamis, N. Doulamis, and C. Loupos, "Deep convolutional neural networks for efficient vision based tunnel inspection," in *2015 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pp. 335–342, Cluj-Napoca, Romania, 2015.

[28] A. Voulodimos, D. Kosmopoulos, G. Veres, H. Grabner, L. Van Gool, and T. Varvarigou, "Online classification of visual tasks for industrial workflow monitoring," *Neural Networks*, vol. 24, no. 8, pp. 852–860, 2011.

[29] A. S. Voulodimos, D. I. Kosmopoulos, N. D. Doulamis, and T. A. Varvarigou, "A top-down event-driven approach for concurrent activity recognition," *Multimedia Tools and Applications*, vol. 69, no. 2, pp. 293–311, 2014.

[30] M. Daszykowski, B. Walczak, and D. L. Massart, "Representative subset selection," *Analytica Chimica Acta*, vol. 468, no. 1, pp. 91–103, 2002.

[31] C. Lalos, A. Voulodimos, A. Doulamis, and T. Varvarigou, "Efficient tracking using a robust motion estimation technique," *Multimedia Tools and Applications*, vol. 69, no. 2, pp. 277–292, 2014.

[32] A. Voulodimos, D. Kosmopoulos, G. Vasileiou et al., "A threefold dataset for activity and workflow recognition in complex industrial environments," *IEEE Multimedia*, vol. 19, no. 3, pp. 42–52, 2012.

[33] D. I. Kosmopoulos, N. D. Doulamis, and A. S. Voulodimos, "Bayesian filter based behavior recognition in workflows allowing for user feedback," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 422–434, 2012.

[34] N. D. Doulamis, A. S. Voulodimos, D. I. Kosmopoulos, and T. A. Varvarigou, "Enhanced human behavior recognition using HMM and evaluative rectification," in *Proceedings of the first ACM international workshop on Analysis and retrieval of tracked events and motion in imagery streams - ARTEMIS '10*, pp. 39–44, Firenze, Italy, 2010.

[35] E. Protopapadakis, K. Makantasis, G. Kopsiaftis, N. Doulamis, and A. Amditis, "Crack identification via user feedback, convolutional neural networks and laser scanners for tunnel infrastructures," in *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pp. 725–734, Rome, Italy, 2016.