Semantic Classification of Monuments' Decoration Materials Using Convolutional Neural Networks: A Case Study for Meteora Byzantine Churches

Nikolaos Bakalos
National Technical University of Athens
Athens, Greece

Sofia Soile
National Technical University of Athens
Athens, Greece

Charalabos Ioannidis
National Technical University of Athens
Athens, Greece

ABSTRACT

https://doi.org/10.1145/3389189.3398000Historic preservation of tangible cultural heritage assets is a process that goes beyond structural integrity to the restoration of the interior decorations, such as wall-paintings or icons since this provides a complete restoration process of the monuments that face both their architectural and functional elements. This process is imperative, as in a lot of cases parts of the assets (e.g., frescoes) are decayed or missing due to the passage of time and other environmental, natural or anthropogenic factors. An indicative paradigm of such a decay is the Byzantine churches in Meteora area, a UNESCO cultural heritage site in Greece. However, the limitations in taking samples from such sights indicate that before such fresco restoration process commences, we first need to semantically classify the monument surfaces into different material types, such as stone, mortar or frescoes. The research challenge imposes this semantic classification process is more evident in cases where the surfaces of the monument are not planar but complex, such as in many byzantine churches carved in rock in Meteora.

In this paper, the semantic classification is achieved using a deep Convolutional Neural Network (CNN) which receives as input two types of data: RGB images of the frescoes to capture textural information and 3D cubes that encapsulate the geometric structure of the surface. RGB images describe visual complexity of the frescoes including texture maps and style. On the other hand, the 3D cubes include triangles of the surface, obtained using photogrammetric methods, describing surface complexity. The CNN consist of two layers; a deep convolutional layer which automatically extracts a set of reliable features from the input raw data and a conventional feedforward neural-based classification layer. To detect the missing items and the material types, overlapped input data are fed as inputs to the CNN as if the network "scan" the decorations to discriminate the type of their materials. The classification performance is tested on real-world destroyed byzantine frescoes of Saint Stephanus Monastery in Meteora.

PETRA '20, June 30-July 3, 2020, Corfu, Greece

ACM ISBN 978-1-4503-7773-7/20/06…$15.00
https://doi.org/10.1145/3389189.3398001

CCS CONCEPTS

•Computing methodologies~Machine learning~Machine learning approaches~Neural networks

KEYWORDS

Semantic Classification, Convolutional Neural networks, Material Detection, Cultural Heritage, Complex Surfaces

1 INTRODUCTION

One salient aspect in historic preservation of monuments is, apart from rehabilitating their structural integrity and retaining the resilience of the building materials, the restoration of the interior decorations, such as murals or icons, since this provides a complete restoration process of the monuments that face both their architectural and functional elements. For instance, in old byzantine churches, such as the ones of the famous monasteries in Meteora area, Greece (UNESCO cultural heritage monuments), one critical issue is to restore missing or decayed parts of wall frescoes which have been destroyed due to several environmental and anthropogenic factors through time. Before such fresco restoration process commences, the semantically classification of the monument surfaces into different material types like stone, mortar or fresco is very helpful. The research challenge of this semantic classification process is more evident in cases, such as in most of the byzantine churches in Meteora, where the surfaces of the monument are not planar but complex (free form surfaces instead of surfaces following a mathematical shape) since the irregular surface of the rocks is incorporated and is part of the church masonry.

In this paper, the semantic classification is achieved using a deep Convolutional Neural Network (CNN) which receives as input two

types of data: (a) RGB images of the walls to capture textural information, and (b) 3D cubes that encapsulate the geometric structure of the surface. RGB images describe visual complexity of the wall including texture maps and style. On the other hand, the 3D cubes include triangles of the surface, obtained using photogrammetric methods, describing surface complexity. The CNN consist of two layers; a deep convolutional layer which automatically extracts a set of reliable features from the input raw data and a conventional feedforward neural-based classification layer. To detect the missing items of the frescoes and the material types, overlapped input data are fed as inputs to the CNN as if the network "scan" the decorations to discriminate the type of their materials. The classification performance is tested on real-world destroyed byzantine frescoes of Saint Stephanus Monastery in Meteora.

1.1 Previous Work

To achieve sustainable protection and restoration of such assets the characterization of building materials and decay is of utmost importance, especially in terms of intervention conservation practices. For the protection of a monument, in most cases, it is forbidden to take samples [1], [2]. Therefore, the scientific community turns to non-invasive and no-contact practices to acquire the necessary information. For the building materials characterization, non-destructive techniques are utilized for the determination of the pathology of a monument. The collection of vast amounts of data using these techniques, can contribute to the protection of cultural heritage assets and also for the decision making on conservation approaches ([1], [2], [3], [4]). Over the last decade, there is an immerse use of digital geometric documentation processes, especially for the creation of three-dimensional (3D) textured models ([5], [6], [7]). The combination of photogrammetric and computer vision algorithms (Structure-from-Motion techniques) may provide accurate and detailed 3D architectural surveys with radiometric information ([8], [9], [10]).

For the protection of a cultural heritage asset, the classification and representation of a monuments' pathology aims to control the decay progress and to improve planning of conservation interventions [1]. Within the framework of cultural heritage assets protection, practical needs emerge regarding the integrated study management and the knowledge deriving from incompatible interventions towards an interdisciplinary integrated approach. Moreover, the academic community adopts such approaches, especially for the investigation of assets construction phases, deformations and restoration practices, for their visualization and projection in multiple digital platforms. The tendency is to use Geographic Information System (GIS) or Building Information Modelling (BIM) software to create 3D models ([11],[12],[13]) incorporating information of the abovementioned disciplines as well as other depending on the project's scope (documentation, visualization, dissemination, restoration etc.)

In cultural heritage the use of machine learning techniques has proved effective in various instances both in static environments (i.e. tangible assets) and even for dynamic environments (intangible cultural assets). In [14] multispectral information and machine learning for non-destructive preservation of cultural assets is used. Moreover, a number of machine learning approaches has been used for semantic classification of cultural heritage assets. In [15] and [16] machine learning techniques and the effect of Industry 4.0 innovations for semantic annotation of cultural heritage assets is studied. In [17] a collection of datasets for benchmarking machine learning techniques in cultural heritage is provided. [18] provides a machine learning framework for the automated classification of heritage buildings. Similar approaches, employing non-machine learning techniques have also been used. Such techniques include thematic mapping ([19]) and hierarchical models ([20]). However, even in the case of user monitoring and time-varying environments deep learning approaches have proved useful, as indicated in [21]. This can be translated into providing semantic information in even intangible cultural heritage assets, as is the case for [22], [23] and [24].

In this work a similar approach based on convolutional neural networks to provide semantic information for identifying decoration materials in tangible cultural heritage assets is presented. This approach was experimentally tested using a dataset captured in a Byzantine church in Meteora area (frescoes in the Katholikon of the Saint Stephanus Monastery).

2 PROBLEM FORMULATION

Let us denote as $y(n) = [P_{1,}P_{2,} \dots P_{N,}]^T$ a N$x1$ vector that contains the probabilities P, that the observations at pixel instance n can be classified as one of N types of materials. Let us now assume that there is a non-linear function that relates probabilities $y(n)$ with some measurable observations $x(n)$. In the following notation, we assume that $x(n)$ are multidimensional tensors of the input data. Assuming a non-linear dependency of the classification output and the previous classification values, we derive a non-linear autoregressive-moving average model:

$$y(n) = g(x(n,) + e(n) \quad (1)$$

where $g(\cdot)$ refers to the non-linear relationship. Vector $e(n)$ is an independent and identically distributed error. Eq. (1) cannot be easily calculated, as $g(\cdot)$ is unknown. The use of machine learning methods can produce an approximation of $g(\cdot)$ in a way that minimizes the error $e(n)$. A feed forward neural network (FNN) can simulate the behavior of such function. However, this FNN model fails at effectively selecting features of high-dimensional space and complex heterogeneous environments. Convolutional Neural Networks have demonstrated excellent representational capabilities in feature selection as in [25]. The proposed filter exploits the effectiveness in feature selection of CNN, in order to select optimal features that enable the classification of the observed behaviors.

After the input layer that receives the current data, we proceed with the convolutional/pooling layers. This layer applies convolutional transformations on the input data so as to maximize the classification performance. The convolutions are executed over the

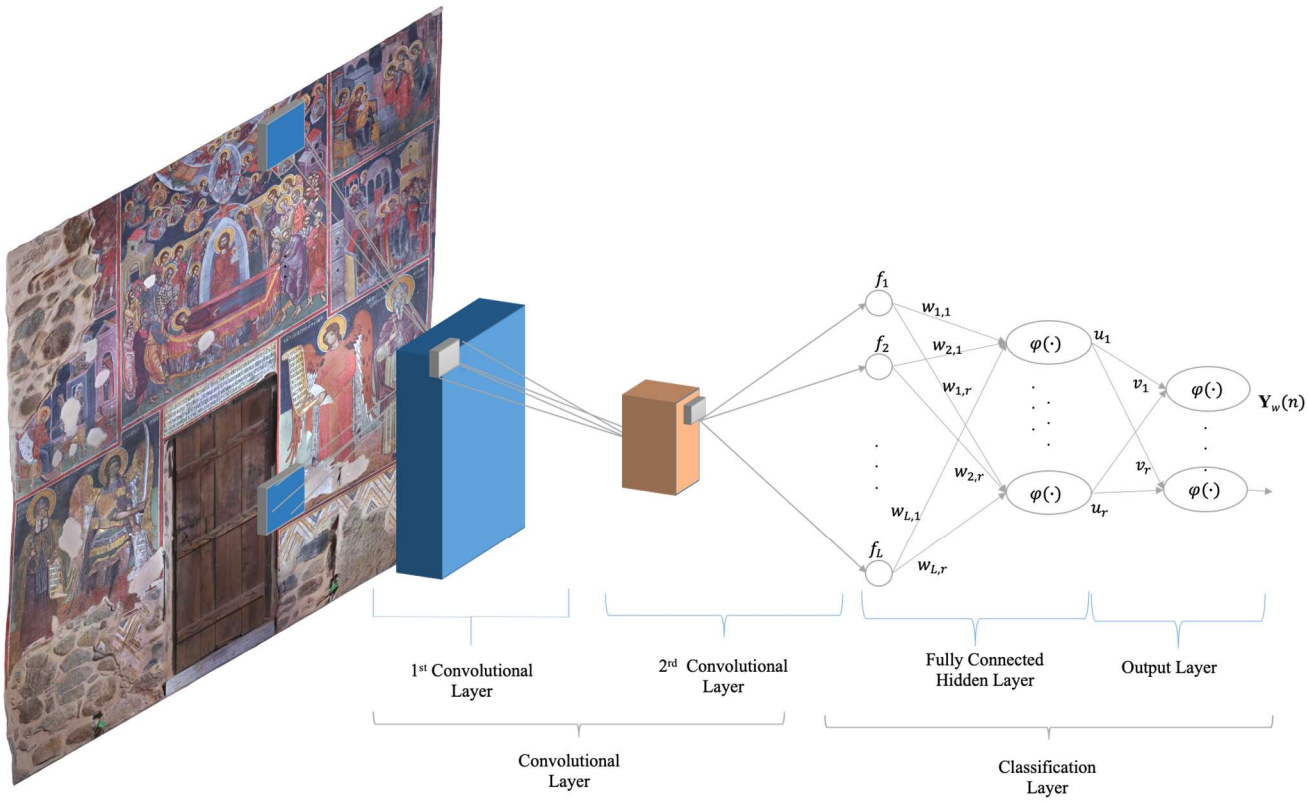

Figure 1: System Architecture

input data and a set of kernels, in order to select appropriate features. The kernel parameters are estimated in a way that minimizes the performance error on a ground-truth training set. The L feature maps, denoted as $f_1, f_2, \ldots, f_L$ are used as inputs in the final (classification) layer. In the experimental evaluation, the convolutional/pooling layers consist of three different convolutional layers, with 5x5x4, 5x5x32 and 5x5x32 respective filter sizes, separated by the ReLU and Max pooling components.

The final component of the filter is the classification layer that receives the $f_1, f_2, \ldots, f_L$ feature maps and triggers a supervised behavior classification. The f_i feature maps are tensors with dimensions that express the spatial attributes and the different modalities of the input data.

The classification layer consists of r neurons, each stimulating a non-linear operation, where the sigmoid is neuron activation function. If we denote as $w_{i,j}$ the weights that connect the i-th feature map f_i with the j-th hidden neuron of the classification layer, then the output of this neuron will be $u_j = \varphi(\mathrm{w}_j^T \cdot f)$, where f is the aggregate feature map concatenating all features f_i and w_j the aggregate weights for the j-th hidden neuron. Then, output will be given as:

$$y_w(n) = \varphi\left(v^T \cdot u\right) \equiv \varphi(z_w(n)) \qquad (2)$$

where u includes all outputs u_j over all the r hidden neurons and v the aggregate weights connecting the r hidden neurons of the classification layer with the output neuron. In Eq. (2), $z_w(n)$ expresses the input of the final output neuron before applying the activation function $\varphi(\cdot)$. In the previous notation, we have assumed that the classification output consists of one neuron. Extension to multiple neurons is straightforward. Subscript w in Eq. (2) denotes the dependence of the classification on the network weights which will be estimated through a learning process. In our configuration, the proposed model consists of 64 hidden layers and two output neurons.

A schematic diagram of the proposed architecture is presented in Figure 1.

3 SYSTEM ARCHITECTURE

In this section the proposed implementation of a Convolutional Neural Network is described.

Convolutional Layer: The purpose of this layer is to apply convolutional transformations on the input data in a way as to maximize classification performance. A set of parameterisable filters (e.g., learnable kernels) is convolved with the input data selecting appropriate features and estimating kernel parameters so

Figure 2: Example sample of the dataset

that performance on a ground truth training set is maximized. The L feature maps, say f_1, f_2, …, f_L, optimally selected by the convolutional layer will be used as input to the final classification layer.

Classification Layer: The Classification Layer receives as input the transformed representations from the convolutional layer, i.e. feature maps $f_1, f_2, \dots, f_L$, and triggers the final (supervised) classification. Normally, feature maps f_i re tensors of a high dimensional grid. The first dimensions express the spatial attributes of the scene, either in 2D or 3D space, while the rest refer to the different modalities (channels) of the input data. In the following, to simplify the notation, we assume, without loss of generality, that the feature maps f_i are scalars. Extension to tensors can be done by exploiting tensor algebra properties and appropriate modification of the inner product operators.

4 PERFORMANCE EVALUATION

4.1 Dataset Description

For the experimental evaluation of the proposed system, data captured from the interior walls of a Byzantine church in Meteora (Katholikon of Saint Stephanus Monastery) are utilised. The walls are covered by frescoes, other painting decorations, stones and mortar.

The RGB image used to capture the textural information is the ortho-image of each wall. This ortho-image is an accurate 2D representation of the textured 3D model on a plane parallel to the mean plane of the (non-planar) wall. The textured 3D model of the interior of the church and the ortho-images of the walls was produced by photogrammetric and SfM techniques. Initially, a large number of RGB images of the interior of the church using a 24mm focal length camera and field measurements of control points for the georeferencing were captured. Then, the dense point cloud representing the DSM of the surfaces, the solid 3D model and, finally, the ortho-images are created.

A part of the ortho-image of the western wall including the entrance of the church is presented in Figure 2. To annotate this image, in collaboration with cultural preservations the DBSCAN [26] algorithm to separate the image in clusters, that are in turn annotated by the preservators, is used. The image and the labels are then fed to the CNN for training.

4.2 Overview of the implementation

The CNN classifier, as well as the other frameworks that are described in Section 4.3, the performance of which was tested against the proposed model were implemented in Python 3.6 using the Keras (1.08) and Tensorflow (2.1.0) libraries. For hyperparameter optimization the Tune library was used. The system was trained on an Intel Core i7-6700K CPU (4GHz) with 2 NVIDIA GTX1080 GPUs.

One of the main drawbacks of deep machine learning techniques is the fact that they require large training datasets, to optimize the algorithm performance and that usually the training of such classifiers is computationally demanding. However, the captured dataset proved enough to reach high levels of accuracy, as is presented in the following sections. Moreover, the training of the CNN required approximately an hour of training time. This is an acceptable computational cost based on the application scenario, though the training was significantly slower than the other testing classifiers.

4.3 Experimental Validation

The classification performance of the proposed Deep Convolutional filter was compared with different classifiers, i.e. (i) a linear kernel SVM, and (ii) an architectures of a traditional feedforward neural network with 2 hidden layers of 10 neurons/layer. The results are presented in Figure 3 and in Table 1. CNN clearly outperforms all the other classifiers. For assessing the performance of the classifiers, we used the popular metrics accuracy (ration of correct predictions over the total of predictions made), precision (number of correct prediction divided by the number of total predictions made), recall (number of correct predictions divided by the total number of elements present in that class) and F1-score (harmonic mean of precision and recall).

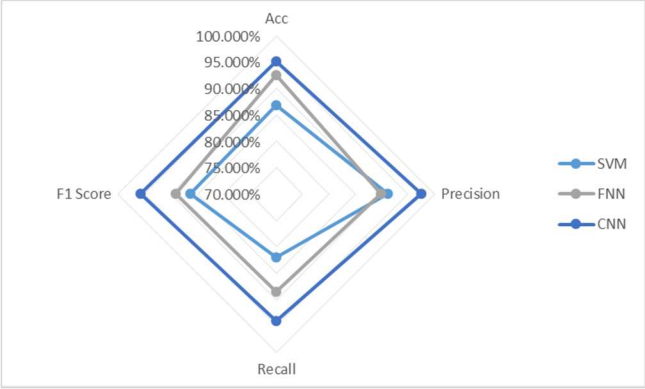

Figure 3: Performance of the various classifiers

Table 1: Classification performance metrics on multimodal experiments

Method	Train Time	Accuracy	Precision	Recall	F1 Score
SVM	33 min	86.78 %	91.10 %	82.08 %	86.35 %
FNN	40 min	92.61 %	89.84 %	88.46 %	89.15 %
CNN	58 min	95.20 %	97.52 %	94.01 %	95.73 %

As is observed the proposed CNN classifier outperforms the other tested methods, while the FNN and SVM classifiers show similar performance, especially when studying the precision of the models. This can be attributed to the ability of the CNN to create appropriate filters that extract representative features that drive the classification stage.

Moreover, the CNN classifier displays not only better accuracy, i.e. has a higher ratio of correctly predicted observations to the total observations. CNN also displays higher precision, and recall.

5 CONCLUSIONS

In this paper we argued the need of an intelligent system that can semantically enrich the 3D models of a cultural heritage asset in a way that facilitates the restoration and preservation of such an asset. We formulate this as a classification problem, were the enriched information takes the form of pixel level annotation over the ortho-image that is extracted from the 3D model of the interior of a Byzantine church in the UNESCO Cultural Heritage Site of Meteora. We utilize a Convolutional Neural Network and demonstrate that this architecture can efficiently extract representative features from the ortho-image and classify the ortho-image pixels over a number of predetermined materials.

ACKNOWLEDGMENTS

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T1EΔK-02859).

REFERENCES

[1] A. Moropoulou, E.T. Delegou, N.P. Avdelidis and A. Athanasiadou. 2005. Integrated diagnostics using advanced in situ measuring technology. In *Proceedings of the 10th DBMC - International Conference on Durability of Building Materials and Components*.

[2] R. Letellier, W. Schmid, F. LeBlanc and F. Recording. 2007. *Documentation & Information Management for the Conservation of Heritage Places - Guiding Principles*. J. Paul Getty Trust: Los Angeles, CA, USA.

[3] P. Salonia and A. Negri. 2003. Historical buildings and their decay: Data recording, analysing and transferring in an ITC environment. *ISPRS Arch.*, vol. 34, 302–306.

[4] J.E. Meroño, A.J. Perea, M.J. Aguilera, and A.M. Laguna. 2015. Recognition of materials and damage on historical buildings using digital image classification. *S. Afr. J. Sci.* 111, 1–9.

[5] G. Guidi, and F. Remondino. 2012. 3D modeling from real data. In Modeling and Simulation in Engineering. Alexandru, C., Ed.; In *Tech Publisher: Rijeka, Croatia*, 69–102. ISBN 978-953-51-0012-6.

[6] T.P. Kersten, and M. Lindstaedt. 2012. Image-based low-cost systems for automatic 3D recording and modelling of archaeological finds and objects. In *EuroMed 2012: Progress in Cultural Heritage Preservation*, Ioannides, M., Fritsch, D., Leissner, J., Davies, R., Remondino, F., Cao, R., Eds.; LNCS vol. 7616, 1-10, Springer: Berlin/Heidelberg, Germany.

[7] F. Remondino. 2014. Photogrammetry—Basic Theory. In *3D Recording and Modelling in Archaeology and Cultural Heritage—Theory and Best Practices*, Remondino, F., Campana, S., Eds., Archaeopress BAR Publication Series 2598, 63-72, Gordon House: Oxford, UK. ISBN 9781407312309.

[8] A. Perez Ramos and G. Robleda Prieto. 2015. 3D virtualization by close range photogrammetry indoor gothic church apses - The case study of church of San Francisco in Betanzos (La Corua, Spain). *ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.,* vol. XL-5/W4, 201–206.

[9] W. Boehler, W G. Heinz and A. Marbs. 2010. The potential of non-contact close range laser scanners for cultural heritage recording. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. 34, 430–436.

[10] A. Georgopoulos, C. Ioannidis, S. Soile, S. Tapeinaki, R. Chliverou, E. Tsilimantou, K. Labropoulos and A. Moropoulou. 2018. The Role of Digital Geometric Documentation for the Rehabilitation of the Tomb of Christ. In *Proceedings of the Digital Heritage 2018 3rd International Congress*

(DigitalHERITAGE), San Francisco, USA. DOI: 10.1109/DigitalHeritage.2018.8810044

[11] M.R. Hess, V. Petrovic, F. Kuester. 2017. Interactive classification of construction materials: Feedback driven framework for annotation and analysis of 3D point clouds. In *Proceedings of the 26th International CIPA Symposium*, vol. IV-2/W2, 343–347, Ottawa, ON, Canada.

[12] D. Costantino, M.G. Angelini. 2015. Three-Dimensional Integrated Survey for Building Investigations. *J. Forensic Sci.*, vol. 60, 1625–1632.

[13] E. Adamopoulos, E. Tsilimantou, V. Keramidas, M. Apostolopoulou, M. Karoglou, S. Tapinaki, C. Ioannidis, A. Georgopoulos and A. Moropoulou. 2017. Multi-sensor documentation of metric and qualitative information of historic stone structures. In *Proceedings of the 26th International CIPA Symposium*, IV-2/W2, 1–8, Ottawa, ON, Canada.

[14] N. Bakalos, N. Doulamis and A. Doulamis, A. 2020. Multispectral Monitoring of Microclimate Conditions for Non-destructive Preservation of Cultural Heritage Assets. *Strategic Innovative Marketing and Tourism*, 641-646.

[15] A.M. Yasser, K. Clawson and C. Bowerman. 2017. Saving cultural heritage with digital make-believe: machine learning and digital techniques to the rescue. In *Proceedings of the Electronic Visualisation and the Arts (EVA 2017)*, BCS Learning & Development Ltd.

[16] C. Mirarchi, A. Pavan, B. di Martino and A. Esposito. 2020. Impact of Industry 4.0 in Architecture and Cultural Heritage: Artificial Intelligence and Semantic Web Technologies to Empower Interoperability and Data Usage. In *Impact of Industry 4.0 on Architecture and Cultural Heritage*, Chapter 13, 306-329, IGI Global.

[17] P. Ristoski, G.K.D. de Vries and H. Paulheim. 2016. A collection of benchmark datasets for systematic evaluations of machine learning on the semantic web. In *International Semantic Web Conference*, 186-194, Springer, Cham.

[18] M. Bassier, M. Vergauwen and B. van Genechten. 2017. Automated classification of heritage buildings for as-built BIM using machine learning techniques. *ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. IV-2/W2, 25-30.

[19] E.S. Malinverni, F. Mariano, F. di Stefano, L. Petetta and F. Onori. 2019. Modelling in HBIM to document materials decay by a thematic mapping to manage the cultural heritage: The case of "Chiesa della Pietà" in Fermo. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XLII-2/W11, 777-784.

[20] L. Xu and X. Wang. 2015. Semantic Description of Cultural Digital Images: Using a Hierarchical Model and Controlled Vocabulary. *D-Lib Magazine*, vol. 21(5/6), DOI: 10.1045/may2015-xu

[21] N. Bakalos, E. Protopapadakis, A. Doulamis and N. Doulamis. 2018. Dance Posture/Steps Classification using 3D Joints from the Kinect Sensors. In *IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, 868-873.

[22] I. Rallis, N. Bakalos, N. Doulamis, A. Doulamis and A. Voulodimos. 2019. Bidirectional long short-term memory networks and sparse hierarchical modeling for scalable educational learning of dance choreographies. *The Visual Computer*, 1-16.

[23] I. Rallis, A. Langis, I. Georgoulas, A. Voulodimos, N. Doulamis and A. Doulamis. 2018. An embodied learning game using kinect and labanotation for analysis and visualization of dance kinesiology. In *10th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games) IEEE*, 1-8.

[24] A. Voulodimos, N. Doulamis, A. Doulamis, A. and I. Rallis. 2018. Kinematics-based extraction of salient 3D human motion data for summarization of choreographic sequences. In *24th International Conference on Pattern Recognition (ICPR)* IEEE, 3013-3018.

[25] N. Bakalos, A. Voulodimos, N. Doulamis, A. Doulamis, A. Ostfeld, E. Salomons, J. Caubet, V. Jimenez, V. and P. Li. 2019. Protecting water infrastructure from cyber and physical threats: Using multimodal data fusion and adaptive deep learning to monitor critical systems. *IEEE Signal Processing Magazine*, vol. 36(2), 36-48.

[26] E. Schubert, J. Sander, M. Ester, H.P. Kriegel and X. Xu. 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, vol. 42(3), 1-21.