frontiers in ICT

ORIGINAL RESEARCH
published: 06 November 2018
doi: 10.3389/fict.2018.00029

Automatic 3D Reconstruction From Unstructured Videos Combining Video Summarization and Structure From Motion

*Anastasios Doulamis**

National Technical University of Athens, Athens, Greece

Social media and collection of large volumes of multimedia data such as images, videos and the accompanying text is of prime importance in today's society. This is stimulated by the power of the humans to communicate with one another. A useful paradigm of exploitation of such a huge amount of multimedia volumes is the 3D reconstruction and modeling of sites, historical cultural cities/regions or objects of interest from the short videos captured by simple users mainly for personal or touristic purposes. The main challenge in this research is the unstructured nature of the videos and the fact that they contain much information which is not related with the object the 3D model we ask for but for personal usage such as humans in front of the objects, weather conditions, etc. In this article, we propose an automatic scheme for 3D modeling/reconstruction of objects of interest by collecting pools of short duration videos that have been captured mainly for touristic purposes. Initially a video summarization algorithm is introduced using a discriminant Principal Component Analysis (d-PCA). The goal of this innovative scheme is to extract the frames so that bunches within each video cluster that contains videos of content referring to the same object present the maximum coherency of image data while content across bunches the minimum one. Experimental results on cultural objects indicate the efficiency of the proposed method to 3D reconstruct assets of interest using an unstructured image content information.

Keywords: 3D modeling, social media, video summarisation, 3D reconstruction, PCA

OPEN ACCESS

Edited by:
Guanghui Wang,
University of Kansas, United States

Reviewed by:
George Livanos,
Technical University of Crete, Greece
Stelios Tsafarakis,
Technical University of Crete, Greece
Konstantinos Tserpes,
Harokopio University, Greece

***Correspondence:**
Anastasios Doulamis
adoulam@cs.ntua.gr

Specialty section:
This article was submitted to Vision Systems Theory, Tools and Applications, a section of the journal Frontiers in ICT

Received: *25 May 2018*
Accepted: *16 October 2018*
Published: *06 November 2018*

Citation:
Doulamis A (2018) Automatic 3D Reconstruction From Unstructured Videos Combining Video Summarization and Structure From Motion. Front. ICT 5:29. doi: 10.3389/fict.2018.00029

INTRODUCTION

Walking in the second decade of 21st century more and more people realize the impact of multimedia and social media in their lives. This is engaged by the rapid increase of internet users which, according to ITU (International Telecommunication Union) statistics, reaches about 3.2 billion users in 2015 (Sanou, 2015). On the other hand, the purchase cost of multimedia capturing devices is going more and more down while most of these are now embedded in laptops, tablets and mobile phones making media production an easy task by everyone, everywhere and in anytime (Wang and Dey, 2013). Finally, the power of the social media and the humans' need to communicate with friends and family, not only by words, texts and chatting but also through the richness of the audio-visual content (Rutkowski and Mandic, 2007), have stimulated new means of interaction with our social surroundings through the usage of social networks like Facebook, Instagram, or Twitter (Soursos and Doulamis, 2012), (Doulamis et al., 2016).

This, in sequel, has boosted the amount, the complexity and the diversity of the digital media being captured, generated, processed, analyzed, and stored across heterogeneous and distributed media repositories and cloud infrastructures such as Picasa, and Flickr (Sevillano et al., 2012). This huge amount of multimedia content, which forms the so-called User Generated Content (UGC) (Li et al., 2018), can be exploited toward a better human-to-human interaction but also for a variety of new application domains in the broad fields of tourism, culture, leisure and entertainment (Kosmopoulos et al., 2009; Kim et al., 2014; Vishnevskaya et al., 2015). For instance, as stated by Ntalianis and Doulamis in (Ntalianis and Doulamis, 2016), the rich media content of the social media can be exploited to create personalized summaries of a human life making him/her *"digitally perpetual"* and leaving his/her mark in the world forever! This means in other words that we can create an album of our activities and lives in space and time which can be used as an historic mark of our family and friends' tree for our descendants to come. Privacy issues should be taken into account in these cases. Only authorized users can access the media content. A detailed categorization of the privacy issues on UGC can be found in (Smith et al., 2012). These issues are out of the scope of this paper but in our case, only freely available data are taken into account.

Another useful usage of this rich multimedia UGC is to be exploited to generate precise three-dimensional (3D) data of our world (Ioannidis et al., 2009). Nowadays, extracting 3D information of the objects and particularly the depth is a process that can be derived either by applying photogrammetric methods from a selected set of images which have been properly captured/generated (Remondino and El-Hakim, 2006) or by using laser scanners (Fritsch and Klein, 2018) or depth sensors such as Time of Flight Cameras (Kim et al., 2009) or Kinect (Nguyen et al., 2012) for static (Guo et al., 2014) and moving objects (Laggis et al., 2017). The main, however, drawback of the photogrammetric approaches that exploit two-dimensional (2D) for the reconstruction is that they need a specific type of cameras to be used for image data capturing or positioning of these cameras on certain orientations with respect to the object of interest so as to get 3D models of high fidelity (Georgousis et al., 2016). This in the sequel implies a high reconstruction cost which is far away from today easy image/video production phases.

An interesting concept is to use the today big multimedia data repositories for the 3D reconstruction phase. This would result in the so-called "wild" 3D modeling in the sense that the image data are from distributed, social or web-based multimedia repositories which have been captured for personal use or other purposes but for sure not for a precise 3D reconstruction (Makantasis et al., 2016). The goal is to exploit unstructured image content to perform a 3D reconstruction scheme by the application of novel content-based filtering methods and visual-based clustering on the use of a spectral scheme. In this paper, we extend the aforementioned concept by focusing on video sequences located on distributed and heterogeneous multimedia platforms. The goal is to exploit the rich visual information of video content to generate 3D models of the scene they depict. We should state that the exploited videos are taken from remote multimedia repositories and these have been generated for personal or business use but not for 3D modeling. Thus, these videos contain a lot of noise and objects of non-interest such as humans in fronts of monuments, moving vehicles, clutter background, etc. In addition, these videos encounter severe camera moving problems since they have been taken without the use of constant tripods and thus the image frames are trembling. To generate the 3D models, initially a video summarization algorithm is applied on the video frames. The objective of a video summarization scheme is to extract a small but meaningful number of key frames from the video sequence able to resemble as much as possible the whole video content (Money and Agius, 2008). The new concept proposed is the use of a discriminant Principal Component Analysis (d-PCA) for summarizing the videos. The d-PCA concept (Wang et al., 2018) was introduced very recently for clustering objects so as to maximize the coherency of foreground against the background. Then, a Structure from Motion (SfM) algorithm is introduced to generate the 3D models.

The proposed concept can be very useful for cultural heritage (CH) applications toward a massive automatic (or at least semi-automatic) documentation of CH objects, a process very useful for their protection and for the implementation of robust resilience actions on them (Yastikli, 2007). More specifically, CH objects, which are not so "attractive and famous" but they are still great in culture and the ancient technology they reveal, often receive inadequate amount of financial support to obtain 3D geocentric models of high fidelity. Furthermore, CH monuments located in poor developing countries or in regions suffering by war, conflicts, or other political uncertainties (e.g., inadequate protection against looting), though the great cultural value they present, cannot attract sufficient financial resources for their accurate documentation (Remondino and Stylianidis, 2016). In all these cases, one can exploit video shots available on the web or in touristic media repositories to provide to the archaeologists/CH experts 3D models of the objects of interest which can be used for their documentation at a very low cost (Doulamis N. et al., 2013; Yiakoumettis et al., 2014).

On the other hand, massive 3D reconstruction can boost augmented reality and virtual reality technology since it will provide a pool of 3D models through which these scientific fields can be evolved (Bruno et al., 2010). Gaming applications, including serious games for education purposes, new fascinating applications to museums' visitors, archaeological tools for documentation and categorization of the objects, or even land monitoring paradigms will be some among other applications scenarios that can be gained by the proposed scheme (Ioannidis et al., 2016).

This paper is organized as follows: A state-of-the art description is given in section Description of the State-of-the-Art and Proposed Contribution. The works described refer to (i) 3D reconstruction modeling approaches, (ii) video summarization, while (iii) the proposed contribution is examined. section Video Parsing and Text-based Filtering introduces the video parsing and text-based filtering method. The new discriminant Principal Component Analysis (d-PCA) algorithm is discussed in section Discriminant Principal

Component Analysis (d-PCA Video Summarization). section On the Fly 3D Reconstruction/Modeling using Structure from Motion shows the on-the-fly 3D reconstruction and modeling method exploiting concepts of Structure from Motion scheme. Experimental results are given in section Experimental Results along with a detailed description of the dataset used and the objective metric applied to judge the efficiency of the video summarization scheme. Finally, conclusions. section Conclusions draws the conclusions.

DESCRIPTION OF THE STATE-OF-THE-ART AND PROPOSED CONTRIBUTION

In this section, we describe the current state-of-the-art in the fields of 3D modeling/reconstruction and video summarization, that is, the two research fields addressed in this paper. Since these fields have been extensively studied in the recent years as is proven by the high number of articles published, we restrict our description in those works that are more relevant to our approach; that is, the ones that present 3D modeling and reconstruction from unstructured video data and video summarization for short term video sequences of cultural or landscape content as the ones encountered in our cases.

3D Modeling/Reconstruction

To derive precise 3D models from a set of cameras, photogrammetric methods should be applied. The first step toward this is to calibrate a set of cameras so as to get precise information on the geometry (Remondino and Clive, 2005). However, camera calibration is not applicable in our case where unstructured visual content is considered, i.e., content available from videos generated for personal (or even for business) use. Then, a set of visual descriptors should be extracted (Rothganger et al., 2006) which should be invariant within any affine transformation. These descriptors can be controlled either from known points (Alsadik et al., 2014) or can be set as a result of an image analysis method (markless description) (Barazzetti et al., 2010; Verykokou et al., 2017). In the following, probabilistic learning or geometry-based analysis or even other classification schemes are applied to reconstruct the depth of the scenery (Gargallo and Sturm, 2005). Nevertheless, this process requires a lot of time. Thus, fast methods have been also introduced to reduce the time while keeping reconstruction accuracy as high as possible (Xia et al., 2013).

A pioneer work that simultaneously solve the camera pose and scene geometry under an automated way is the Structure from Motion (SfM). The method exploits the bundle adjustment technique based on matching features into multiple overlapping images (Bolles et al., 1987; Westoby et al., 2012). This method has been extended to modeling non-rigid structures, i.e., modeling of the shape of objects which are deformable. The so-called Non-rigid structure from motion (NRSfM) recovers the shape and the pose of an object which is deforming in time from a set of monocular cameras (Torresani et al., 2008).

A few works have been proposed for handling the problem of unstructured image data as we address in this paper. More specifically, Dorninger and Nothegger (2007) applies a 3D segmentation for unstructured point clouds. The results have been applied for modeling of buildings, an important task in photogrammetry and remote sensing. The work of (Makantasis et al., 2016) finds a set of relevant images located on distributed and heterogeneous media repositories to derive a precise 3D reconstruction. The results target tangible cultural heritage objects such as monuments, historic regions and buildings. 3D reconstruction from multi-view unstructured images is also proposed in (Zhang and Chen, 2014). The approach analyses 3D planar primitives refined by the RANSAC algorithm (Schnabel et al., 2007) and then the adjacent regions of the planar primitives are estimated to find 3D intersection lines on the respective faces.

Recently, the unstructured 3D modeling has been extended to include the time dimension. In this case, the analysis focuses on the creation of precise 4D models (3D geometry plus the time). The works proposed in this area either use a Bayesian approach for the analysis (Huang et al., 2016) or localize similarities on the image data to accelerate the reconstruction process through time (Doulamis A. et al., 2013; Ioannides et al., 2013).

Video Summarization

Some techniques for video summarization exploits temporal variations of feature vector trajectory to find out characteristic points on the content through which the key frames are extracted. The key idea of these approaches is to localize on the fluctuation of the feature trajectory salient points such as peaks or curvatures and then to extract the key frames at the time instances of the salient points (Doulamis et al., 2000b; Kuanar et al., 2015; Kim et al., 2016). The main advantage of these approaches in video summarization domain is the fact that they can discriminate periodic content patterns and thus to differently handle two similar scenes when these are posted at different time intervals in the video sequence. Although such a property is seen as an advantage for abstracting video sequences, it is a drawback in our case in which video summarization triggers the extraction of a set of characteristic frames through which 3D reconstruction/modeling will be accomplished. In the same framework, the works of (Panagiotakis et al., 2007, 2009) applies an Iso-Content analysis to localize the key frames as the ones placed on the "same" (iso) content distance in the sequence. Other techniques extract a short video summary instead of key frames to present an abstract form of video sequences (Cernekova et al., 2006; Mademlis et al., 2016).

Some other video summarization algorithms select the most discriminant frames in terms of visual content as the key ones. More specifically, a graph-based clustering method for video summarization is presented in (Ngo et al., 2005) while the use of Delaunay clustering is proposed in (Mundur et al., 2006). Min-max optimization framework is introduced in (Li et al., 2005) while a hyper-graph clustering is recently presented in (Ji et al., 2018). A stochastic algorithm that extracts the most representative key frames by minimizing the cross-correlation criterion is introduced in (Avrithis et al., 1999). The same work was improved under a fuzzy framework in (Doulamis et al.,

2000a) and extended to stereoscopic video sequences in which two stereo pairs of each video frame are available in (Doulamis et al., 2000). Finally, in (Meng et al., 2018) a multiview-based video summarization is presented through representative selection.

The Proposed Contribution

This paper proposes a new 3D reconstruction and modeling algorithm that exploits short video sequences generated by simple users mainly for personal use (User Generated Content). We assume that the short videos depict the same scene of a scenery, a building or a monument the 3D model of which we need to construct. The short UGC videos are parsed from multimedia repositories. Initially, a text-based filtering is proposed as in (Makantasis et al., 2016) to refine the short videos with respect to their captions generating a pool of videos each showing the same scene, building or monument. This is applied to refine video sequences so as to improve the reconstruction analysis at the later stages. Only relevant video shots will be taken into account. Then, the collected pool of short videos is summarized by the application of a discriminant algorithm. In our case, we propose a novel video summarization scheme that is based on a discriminant Principal Component Analysis (d-PCA) as presented in the very recent work in (Wang et al., 2018). Discriminant Principal Analysis has its goal to extract the most significant information from one dataset, that is, the prominent information.

The method proposed in (Wang et al., 2018) was applied to recognize handwritten digits and frog images. In this paper, we properly modified and extend this approach to extract a small but meaningful number of key frames from a pool of short videos depicting the same scenery. More specifically, we initially construct a visual feature vector by extracting ORB descriptors from each video frame. ORB can identify salient parts in image content being invariant under affine transformations. Then, we modified the d-PCA to be applicable to time series as video sequences are instead of data collections as the original d-PCA algorithm is applied to. We also introduce the concept of bunches within each cluster so as to differentiate image frames with respect to their angles and orientations. Finally, we propose a modification of the scheme as regards the optimal key frame selection in time. Having extracted the key frames from the videos, we then assume that these can represent as much as possible the whole video content and provide an adequate information for 3D modeling and reconstruction.

The reason we select the d-PCA method for video summarization instead of other techniques proposed in the literature is due to (i) the nature of our video sequences and (ii) the final objective we have, i.e., to derive 3D models of objects from video shots being captured for different purposes than 3D reconstruction (e.g., for touristic ones). The first fact implies that our video shots are of short duration, usually captured the same object (e.g., a monument) of interest under different angles and scale. The second fact means that we need to identify a sufficient number of views of the object to get a detailed 3D reconstruction while simultaneously to "get rid off" of views that contribute a little to the reconstruction process. The d-PCA is designed to find clusters on visual data so that (i) content across clusters is as different as possible while (ii) the content within a cluster is as coherent as possible. The first criterion leads to a selection of the minimally required number of clusters (i.e., views) that we need to take into account for a 3D reconstruction. The second criterion leads to a selection of the most representative object views among a plethora of similar ones being captured on video shots. This constitutes the reason we adopt d-PCA for video summarization rather than other techniques which have been mainly designed to summarize long duration videos consisting of different scenes, totally different content and with the purpose to give a quick overview of the content of the video sequence instead of detecting different object views to derive 3D models.

The selected key frames are then fed as inputs to a SfM algorithm for performing the reconstruction. Since only a small number of representative frames is used as inputs to the SfM, the time required for modeling is optimized, while simultaneously we keep 3D reconstruction accuracy as high as possible. **Figure 1** shows a block diagram of the proposed architecture and the main steps proposed to derived an on-the-fly 3D modeling and reconstruction from unstructured UGC short videos distributed over heterogeneous multimedia repositories.

As we can see from **Figure 1**, the proposed scheme consists of:

(a) A video parser: which is responsible to localize the videos from the distributed multimedia repositories or on social media platform. A set of N repositories/social platforms are considered.
(b) Text-based filtering: The goal is to refine the parsed videos so as to group together the ones that present similar textual description, e.g., similar captions. The idea is to perform a kind of semantic filtering on the data by clustering together videos that depict the same scenery, building, monument, region, etc. It is clear that many outliers will be encountered due to inconsistency of the text in describing semantic meanings.
(c) A d-PCA-based Video Summarization: This component implements the new video summarization algorithm proposed in this paper which is based on a discriminant Principal Component Analysis (d_-PCA). The main objective is to extract a set of representative frames from the pool of short videos depicted similar visual content.
(d) SfM algorithm: We then proceed with the application of the Structure from Motion (SfM) algorithm through which the 3D reconstruction/modeling is accomplished.

Overall, the proposed technique aims at providing 3D reconstruction models for objects of interest (especially cultural ones) from non-structured User Generated Content (UGC) which has been captured for different purposes (e.g., touristic) than 3D modeling. The main goal is to achieve a massive 3D reconstruction of objects and monuments of interest rather than using high cost photogrammetric methods. Thus, we exploit on the fly 3D reconstruction by analyzing video content from short duration video sequences. Instead, the approaches presented in the literature in the field are focused on accomplishing fidelity 3D models or

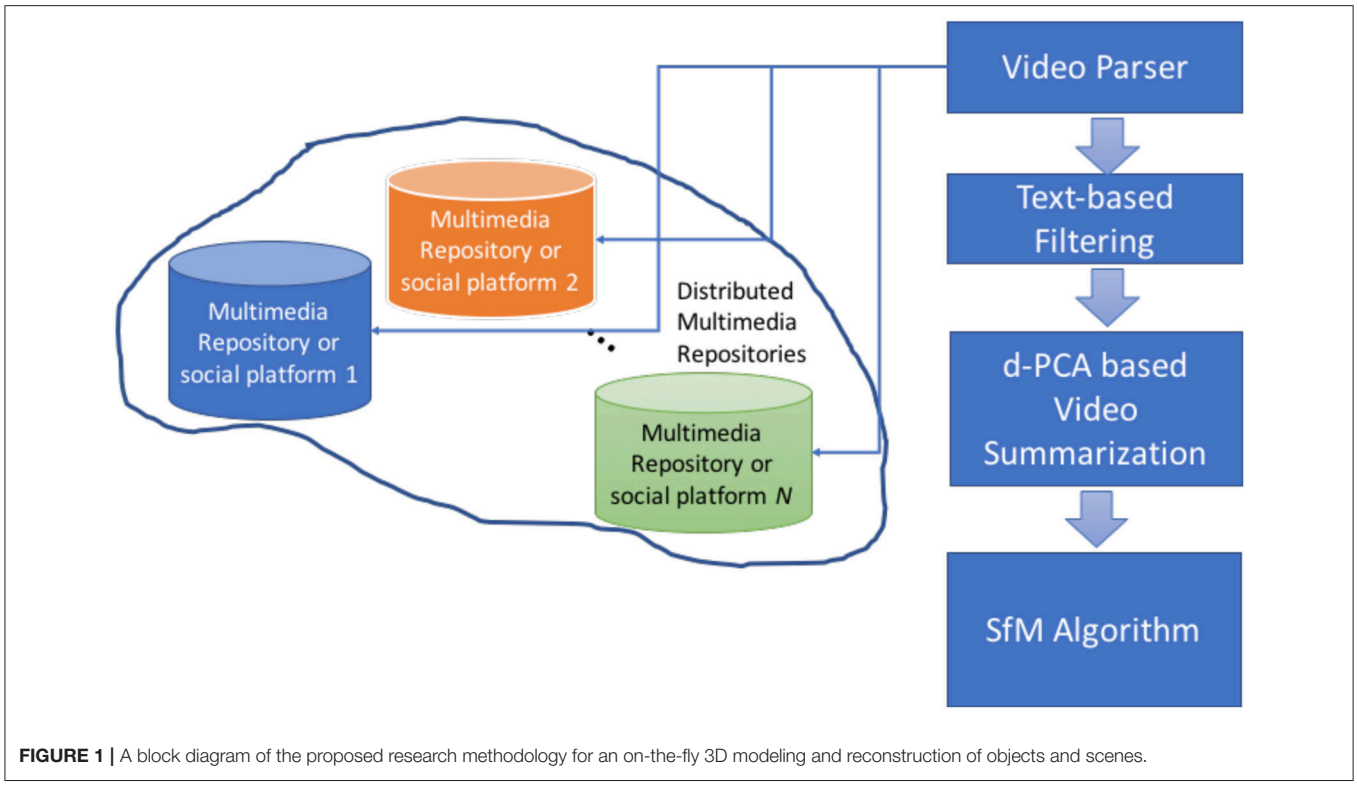

FIGURE 1 | A block diagram of the proposed research methodology for an on-the-fly 3D modeling and reconstruction of objects and scenes.

on summarizing video sequences in the sense of automatic extracting small short duration trailers. Thus, this paper targets a challenging issue in 3D computer vision society; how to accelerate 3D reconstruction and achieve a massive 3D modeling of objects of interest by exploiting existing video content which has been captured for purposes different than 3D analysis.

VIDEO PARSING AND TEXT-BASED FILTERING

The two components described in this section are (i) the video parsing and (ii) the text-based filtering of the video content.

The algorithm deployed detects videos from distributed and heterogeneous repositories and/or platforms of social media. The idea is to distinguish the videos from other multimedia sources such as images, audios, sounds, graphics and texts. This way, the algorithm localizes only video sources. The suffices of the data are used as a parser filter to get the videos. File suffices that correspond to compressed or uncompressed video files are used to filter out than the remaining ones. A secure framework is adopted for the multimedia parsing as the one proposed in one of our previous works (Halkos et al., 2009). The objective is to allow for the parser to complete the search without the need to download the multimedia content beforehand and without the content providers (i.e., the multimedia repositories or the owners of the social media) to need to buy the parsing technology. This way, we retrieve V_i short videos from distributed repositories and/or social media platforms. Each video sequence can be considered as a set of $V_i = \{\cdots f_{i,j} \cdots\}$ where $f_{i,j}$ denotes the j-th frame of the i-th video sequence.

Regarding the text-based filtering, the captions or text descriptions of the videos are parsed. Then, a simple linguistic analysis is accomplished to take into account word similarities. Videos that fall into the same textual groups are clustered together to form pools of videos that share similar content. In other words, we form video clusters that share similar textual semantics in terms of the content they represent. Let us denote as $C_k = \{V_i : i \in \tau_k\}$ where τ_k refers to the k-th similar descriptions derived from the text-based filtering. Due to textual inconsistency and erroneous descriptions the number of outliers in these videos may be large. That is, several $V_i \in C_k$ may depict visual content quite different than the respective text-based semantic description τ_k. For instance, let us assume that one cluster C_k gathers videos the textual description of which is aligned to "The Parthenon." In this cluster, videos captured from taverns named "The Parthenon" can be also collected. The content, however, of these video is not in compliance with the monument "The Parthenon." To improve clustering accuracy, we proceed with a geo-tag restriction. That is, the members of a cluster are further decomposed into geospatial clusters where this information is available. Videos of erroneous or misleading descriptions that correspond to different geo-tags are removed from the respective cluster improving its coherency.

For the removal of the outliers, the visual content of each cluster is spanned into multi-dimensional manifolds taking as inputs invariant visual descriptors such as the Oriented Rotated Brief (Rublee et al., 2011) as adopted in (Makantasis et al., 2016). In the following, a dense–based clustering algorithm is applied

such as OPTICS (Ankerst et al., 1999) to remove the outliers and retain only the most concrete videos within each cluster in terms of visual content description.

DISCRIMINANT PRINCIPAL COMPONENT ANALYSIS (D-PCA VIDEO SUMMARIZATION)

A novel video summarization method is adopted in this paper for key frame selection. The method is derived from the recent article in (Wang et al., 2018) applying, however, for recognition of handwritten digits. The goal is to look inside each video cluster, say the C_k, and form representative bunches (sub-clusters) within each cluster so that (i) the elements within each bunch (intra-bunch) to share maximum coherency in terms of visual similarity, while (ii) the elements across bunches (inter-bunch) to be as far as possible in terms of visual similarity.

Let us denote as $f_{i,j}^{(k,l)}$ the j-th frame of the i-th video $V_i \in C_k$ and let us assume that this frame belongs to the l-th bunch creating within the cluster C_k. We denote in the following this bunch as $B^{(k,l)}$. For each video frame, visual descriptors are extracted to better represent its video content. Let us denote these descriptors as $\mathbf{d}_{i,j}^{(k,l)}$. The ORB visual descriptor is extracted in our case to form the vector $\mathbf{d}_{i,j}^{(k,l)}$. To clarify our notation, let us denote as $\mathbf{b}_i^{(k,l)}$ one vector element of the bunch $B^{(k,l)}$. Then, we can create the covariance matrix for all elements of the bunch $\mathbf{b}_i^{(k,l)}$ as

$$\mathbf{C}_{intra} := \frac{1}{\left|B^{(k,l)}\right|} \sum\nolimits_{for\ all\ i} \mathbf{b}_i^{(k,l)} \cdot \left(\mathbf{b}_i^{(k,l)}\right)^{\mathrm{T}} \tag{1}$$

and the covariance matrix across the elements of two bunches

$$\mathbf{C}_{inter} := \frac{1}{\left|B^{(k,l)}\right|} \sum\nolimits_{for\ all\ i\ and\ l \neq m} \mathbf{b}_i^{(k,l)} \cdot \left(\mathbf{b}_i^{(k,m)}\right)^{\mathrm{T}} \tag{2}$$

Then, the goal is to find a vector, say $\mathbf{u}$, such that

$$max_{[\mathbf{u}]_2=1} \frac{\mathbf{u}^{\mathbf{T}} \cdot \mathbf{C}_{intra} \cdot \mathbf{u}}{\mathbf{u}^{\mathbf{T}} \cdot \mathbf{C}_{inter} \cdot \mathbf{u}} \tag{3}$$

Equation (3) means that we should extract video frames the visual content of which as is being represented by the PRB descriptor, should be "present" (similar) in the relevant bunch and not being present (dissimilar) in the "background data" that is, in the other bunches. This mathematical formulation is relevant to the discriminant PCA as proposed in (Wang et al., 2018). However, in this paper we have properly modified the d-PCA notation to be relevant for a video summarization case.

Problem Solution

Generally, matrix $\mathbf{C}_{inter}$ is full rank since the collected videos have not been captured under exactly the same conditions. Thus, it can be eigen-decomposed as

$$\mathbf{C}_{inter} = \mathbf{U}_{inter}^{\mathrm{T}} . \Lambda_{inter} \cdot \mathbf{U}_{inter} \tag{4}$$

where matrices $\mathbf{U}_{inter}$ and Λ_{inter} refers to the eigen-vectors and values of the covariance matrix $\mathbf{C}_{inter}$. If we define the

$$\mathbf{C}_{inter}^{1/2} = \sqrt{\mathbf{C}_{inter}} = \Lambda_{inter}^{1/2} \cdot \mathbf{U}_{inter} \tag{5}$$

and set as a new variable , then the solution of Equation. (3) can be obtained as

$$\mathbf{u}^* = \mathbf{C}_{inter}^{1/2} \cdot \mathbf{v}^* \tag{6}$$

In Equation (6), $\mathbf{u}^*$ and $\mathbf{v}^*$ are the optimal vectors of $\mathbf{u}$ and $\mathbf{v}$ respectively.

Leveraging Lagrangian duality as in (Wang et al., 2018), the optimal solution of Equation (3) can be given as the right eigenvector of the matrix . This can be proven since Equation (3) can be re-written as

$$max_{\mathbf{u}} \quad \mathbf{u}^{\mathbf{T}} \cdot \mathbf{C}_{intra} \cdot \mathbf{u} \tag{7a}$$

$$\text{subject to } \mathbf{u}^{\mathbf{T}} \cdot \mathbf{C}_{inter} \cdot \mathbf{u} = b \tag{7b}$$

which is fact a Lagrange multiplier problem. The solution of Equation (7) is valid for some constant $b>0$ which is set such that $\mathbf{u}_2 = 1$. One possible solution (7b) is to set b=1 and normalize the solution of (7). Equation (7) can be re-written as a Langrage multiplier problem as

$$\mathcal{L}(\mathbf{u}; \lambda) = \mathbf{u}^{\mathbf{T}} \cdot \mathbf{C}_{intra} \cdot \mathbf{u} + \lambda \cdot (1 - \mathbf{u}^T \cdot C_{inter} \cdot \mathbf{u}) \tag{8}$$

To solve the optimization imposed by Equation (8), we exploit notions from the generalized eigen-value problem. That is, it holds that

$$\mathbf{C}_{intra} \cdot \mathbf{u}^* = \lambda \cdot \mathbf{C}_{inter} \cdot \mathbf{u}^* \tag{9a}$$

or equivalently it holds that

$$\mathbf{C}_{inter}^{-1} \cdot \mathbf{C}_{intra} \cdot \mathbf{u}^* = \lambda \cdot \mathbf{u}^* \tag{9b}$$

Equation (9) implies that the optimal solution $\mathbf{u}^*$ is the eigenvector of matrix x. By integrating the constraint of (7b) into (7a) we can derive that

$$\left(\mathbf{u}^*\right)^{\mathrm{T}} \cdot \mathbf{C}_{intra} \cdot \mathbf{u}^* = \lambda^* \cdot \left(\mathbf{u}^*\right)^{\mathrm{T}} \cdot \mathbf{C}_{inter} \cdot \mathbf{u}^* = \lambda^* \tag{10}$$

From Equation (10), it is clear that the optimal solution of (7) is given as the largest eigenvalue of the matrix $\mathbf{C}_{inter}^{-1} \cdot \mathbf{C}_{intra}$.

Extracting Key Frames

Having estimated the optimal vector $\mathbf{u}^*$, we can then proceed with the identification of the key frames within each video cluster C_k and bunch $B^{(k,l)}$. In particular, the optimal vector $\mathbf{u}^*$ contains the indices of frames $f_{i,j}^{(k,l)}$ that should be assigned to the l-th bunch of the k-th cluster. This way, the bunches contain almost similar frames in terms of visual content they represent. In addition, content coherency across bunches of the same video cluster is minimal. The most representative frame is chosen as the one that is closest to the centroid of the bunch. That is,

$$\mathbf{m}^{(k,l)} = \sum\nolimits_{for\ \ all\ i \epsilon B^{(k,l)}} \mathbf{b}_i^{(k,l)} \tag{11a}$$

$$f_{i^*,j}^{(k,l)} = \underset{for\ \ all\ i,j}{\text{argmin}}\ d(\mathbf{d}_{i,j}^{(k,l)}, \mathbf{m}^{(k,l)}) \tag{11b}$$

where $f_{i^*,j}^{(k,l)}$ is the key frame (index i^*) of the j-th video belonging to the l-th bunch of the k-th semantic cluster. In Equation (11b), function refers to the distance between two feature vectors, the one containing the descriptors of the frames $\mathbf{d}_{i,j}^{(k,l)}$ and the mean feature vector of the respective bunch $\mathbf{m}^{(k,l)}$.

In case that more key frames should be extracted per bunch, the more uncorrelated among them are selected as in (Doulamis et al., 2000a). The goal is to find out the furthest frames in terms of visual content representation and depict them as the more representative ones.

ON THE FLY 3D RECONSTRUCTION/MODELING USING STRUCTURE FROM MOTION

The extracted video frames are fed as inputs to a Structure from Motion (SfM) component through which the 3D reconstruction and modeling is accomplished. The main difference of SfM than conventional photogrammetric methods is that the geometry of the scene, the position of the cameras and the orientation is solved automatically without the need of the knowledge of the targets. The latter in conventional photogrammetric methods should be a prior known. To derive the automatic solution of the aforementioned features, SfM exploits an iterative method which is known as bundle adjustment procedure (Triggs et al., 1999). This procedure exploits the visual descriptors as derived from the aforementioned stage, and the selected key frames of section Discriminant Principal Component Analysis (d-PCA Video Summarization) that form a set of overlapping images on the scenery user generated content we want to reconstruct. The extracted visual features should be invariant in scaling rotation and in general under any affine transformation, while they should be robust to illumination changes. Thus, scaled-based features should be identified.

In particular, in SfM the 3D position and location of the camera and the 3D location of the control points are not a priori known. The position of the camera and the scene geometry are automatically reconstructed through the automatic identification of matching features across a set of multiple cameras. Since the scale and orientation are given under relative coordinates a small number of known ground control points (GCPs) should be provided to transform the relative coordinates to absolute coordinates (Westoby et al., 2012).

The first stage of the SfM is to extract a set of reliable points on the images. In this paper, the ORB visual features are extracted (Rublee et al., 2011) since they provide higher accuracy, robustness under affine transformations and illumination fluctuations and they are simultaneously higher executed than other conventional visual descriptors like SIFT or BRIEF. On the extracted visual keypoints, the sparse bundle adjustment method (Triggs et al., 1999) is applied to estimate the position of the camera and the point-cloud. The latter is of low density, that is, a sparse point cloud is generated. For the matching, density-based clustering schemes are employed served on image multi-dimensional manifolds. This way, moving objects such as humans in a scene are automatically removed and the tangible background content is captured for 3D modeling and reconstruction. In the following, a similarity transformation is exploited to reconstruct the camera position from the key point correspondences followed by triangulation through which the 3D point positions are estimated and the whole geometry is reconstructed. To increase the density of the sparse derived point cloud, dense-based algorithms are applied such as the semi-global algorithm (Hirschmüller, 2008).

EXPERIMENTAL RESULTS

In this section, the experiments conducted are analyzed and some results are depicted to demonstrate the efficiency of the proposed scheme. Section Dataset Description describes the dataset used while section Objective Criteria discusses on objective criteria and metrics that are used while section Experiments shows the experimental results.

Dataset Description

The dataset used is a collection of 5,732 videos that have been gathered from multimedia repositories distributed located over the Web and the Twitter. The latter is a social medium that allows users to chat through short messages while links on images and videos can be also posted. The collection of the videos from the distributed multimedia platforms have been performed by the search tool of (Ioannides et al., 2013) under the framework of 4D-CH-World project (Doulamis et al., 2018). From the Twitter, the respective API has been used to gather the videos as described in (Doulamis et al., 2016). All videos are user generated, they are of very short duration (from few seconds up to several minutes) and they usually depict some cultural sites, monuments or buildings of interest since they have been captured for touristic use of simple users. Thus, the captured visual content suffers from high resolution analysis, specific orientation of the cameras, lack of image content overlapping for some regions especially the ones which are not so accessible by the simple users. In addition, in some of them, the camera content is "flickering" due to the hand movements of the users.

The main challenge of these video dataset is that the content is often "contaminated" with information which is not relevant to the sites we want the 3D reconstruct. For example, humans are often present in front of a monument to verify their present in the place. This noise of the data is removed in our case by two ways. The first is through a projection of the video content received upon the respective textual descriptions they accompany them. Videos whose captions or textual descriptions are not in aligned with the average content information are removed. In addition, we further refine the video content by removing outliers. This is done as described in section Video Parsing and Text-based Filtering by (a) applying the ORB visual features (Ankerst et al., 1999; Rublee et al.,

2011) on each video frame, (b) placing the keypoints onto a multidimensional manifold and (c) then deploying the OPTICS dense-based clustering algorithm (Ji et al., 2018) to remove the outliers.

The content of the datasets has been annotated by domain experts. The annotation is performed at two categories; the one that includes all frames that belong to a site class and the other the frames that belong to the outliers' images. On this content, we create the ground truth dataset used to evaluate the results. In addition, the domain experts annotate the content with respect to the potential geometric views we need to get a complete 3D reconstruction. This way, we can evaluate the required number of frames within each bunch of a cluster so as to proceed with an assessment of the coherency of the bunches created.

Python platform is used for extracting the ORB descriptors and the d-PCA-based video summarization algorithm. In addition, the Natural Language toolkit of Python is exploited to count the words' frequencies for all the retrieved tweets. For the 3D reconstruction, we exploit SfM scheme as being provided by the PhotoScan Agio 3D reconstruction platform. Other 3D reconstruction tools such as the open mic mac cab ne also exploited (Verykokou et al., 2017).

Objective Criteria

Objective criteria are used to assess the video summarization approach with and creation of the video clusters and cluster bunches. For the evaluation, the aforementioned dataset is exploited. The two criteria adopted are the Precision which is measured as

$$P = \frac{|S_{re}|}{|S_{su}|} = \frac{|S_{gt} \cap S_{su}|}{|S_{su}|} \tag{12}$$

where S_{re} is a set that contains the relevant image data, i.e., the intersection of the data belonging to the ground truth over the ones getting by the proposed summarization algorithm and the Recall defined as

$$R = \frac{|S_{re}|}{|S_{gt}|} = \frac{|S_{gt} \cap S_{su}|}{|S_{gt}|} \tag{13}$$

Precision actually measures the percentage of the data that have been correctly clustered over the total ones while recall the percentage of the data correctly clustered over the ground truth ones. That is, the two criteria play the role of true positives and true negatives. By combining the two criteria, we can have that

$$F1 = 2 \cdot \frac{P * R}{P + R} \tag{14}$$

$F1$-score actually compensates the two aforementioned criteria of Precision and Recall.

Experiments

Video Summarization Performance

The ground truth dataset described above which has been annotated by domain experts is exploited in this paper to verify the efficiency of video summarization results. The performance is measures using the objectives criteria of Precision, Recall and $F1$-Score as discussed in section Objective Criteria. **Table 1** shows the average results obtained when the d-PCA video summarization algorithm is applied. In the same table, we also depict some comparisons of the proposed methods with two other video summarization techniques. The first of the compared ones adopts a minimization of a cross-correlation criterion to perform the summarization. This way, the most un-correlated video frames are selected as the most suitable ones. The second compared method belongs to the category of techniques that exploit the temporal variation of the feature vector trajectory to perform the analysis. In this case, the results are even lower than the first approach.

The compared results indicate that our d-PCA scheme for video summarization is more suitable in our case where we need to derive 3D reconstruction models from short duration videos than other traditional video summarization algorithms. This means that d-PCA can better select a great number of object views of different angles and orientations (precision value) while simultaneously selects all potential views needed for reconstruction process (recall values). On the other hand, other traditional state-of-the-art video summarization algorithms are better for detecting frames that are mostly uncorrelated [e.g., the algorithm in Panagiotakis et al. (2009)] or presents peak variations in the feature space [e.g., the algorithm in Torresani et al. (2008)].

We have chosen these two methods for video summarization to be used for comparisons with the proposed d-PCA approach due to the fact that they cover the whole range of video summarization methods by detecting (i) content which is visually irrelevant [uncorrelated-see the (Avrithis et al., 1999) approach] or by detecting (ii) periodic motion patterns [the (Doulamis et al., 2000a) approach]. These two approaches represent the whole framework of a video summarization scheme. Regarding the time efficiency of these methods, the work of (Doulamis et al., 2000a) can be implemented in real-time and is suitable even for consumer electronics devices. Our d-PCA approach is more adequate for finding different views and orientations of an objects and thus for 3D reconstruction. In this case, the time needed for the reconstruction can be greater than real-time since the goal is not to extract a trailer of a video sequence in short time but to minimize the time required for the reconstruction by discarding similar object views.

TABLE 1 | Precision, Recall, and F1-score results for the proposed d-PCA video summarization algorithm and comparisons with other methods.

Video summarization	Precision	Recall	*F1*-score
d-PCA	0.78	0.72	0.748
Cross correlation (Avrithis et al., 1999)	0.72	0.68	0.70
Feature vector fluctuation (Doulamis et al., 2000a)	0.67	0.65	0.66

3D Reconstruction Efficiency

We depict experimental results to show the efficiency of 3D reconstruction when the proposed video summarization algorithm has been adopted. The data content mostly refers to cultural heritage monuments in which a great amount of distributed multimedia content is available. In particular, in **Figure 2A**, we depict the reconstruction accuracy using the SfM scheme when 100 images have been selected from a pool of videos, the content of which illustrates visual information from the Monument to the Discoveries (Padrão dos Descobrimentos) in Lisbon, Portugal. This monument was selected since it contains many geometric details, like the structure of the humans (seamen) on the boat. The voids detected in the reconstruction are due to the fact that the available image content does not contain adequate information to properly reconstruct the monument. **Figures 2B–D** presents the 3D reconstruction when a smaller number of image frames is selected while simultaneously we consider the remaining image data are outliers. The partition is made randomly so that some salient images of the monument may be lost. This is the main reason in which the reconstruction accuracy of the monument significantly deteriorates as the number of image frames used as inliers drops. In particular, when the number of inliers is the 60% of the initial ones the 3D reconstruction performance is so bad that the monument is not even recognizable. The main results of such dramatic deterioration of the 3D reconstruction accuracy is due to the position of the Monuments to the Discoveries. Its position on the edge of the sea does not allow a complete, spherical monitoring of the content for all angles and orientations and the performance of the SfM cannot guarantee a sufficient reconstruction accuracy.

We need to stress here that the best reconstruction achieved in **Figure 2A** uses only a small limited number of image frames than the number usually used for a SfM. This means in other words that, although a small number of images frames is actually exploited the reconstruction results is of relatively sufficient quality. Another interesting point is that the selected images fed as inputs to the SfM are not suitable selected to reconstruct the whole geometry. Instead, they have been extracted using the proposed d-PCA video summarization algorithm. This notion proves the significance of our scheme. Using only a very small dataset of images being captured for totally different reasons than ours a sufficient on-the-fly 3D reconstruction of sites of interest is achieved.

Similar results are noticed for another prominent monument, the Eiffel Tower in Paris, France. The results in **Figure 3** start with the extraction of 100 image frames from a pool of short videos that depict the Eiffel Tower monument. Again, we note that a sufficient reconstruction is achieved when a small and unstructured number of frames is selected. If the number of frames is reduced and the outliers is simultaneously increased the accuracy is deteriorated as well but, in this case, it keeps in sufficient levels of details. This is due to the position of the number which allow the video shooting from all its potential angles and orientations. This is not case for the 3D reconstruction of **Figure 2**.

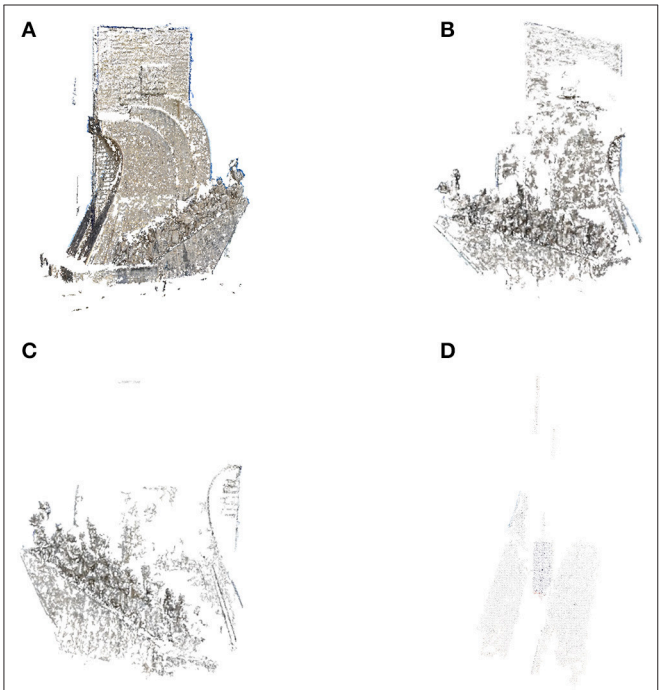

FIGURE 2 | The effect of the number of the outliers in reconstruction performance using the SfM algorithm. **(A)** 100 inliers, 0 outliers. **(B)** 90 inliers, 10 outliers. **(C)** 70 inliers, 30 outliers. **(D)** 60 inliers, 40 outliers.

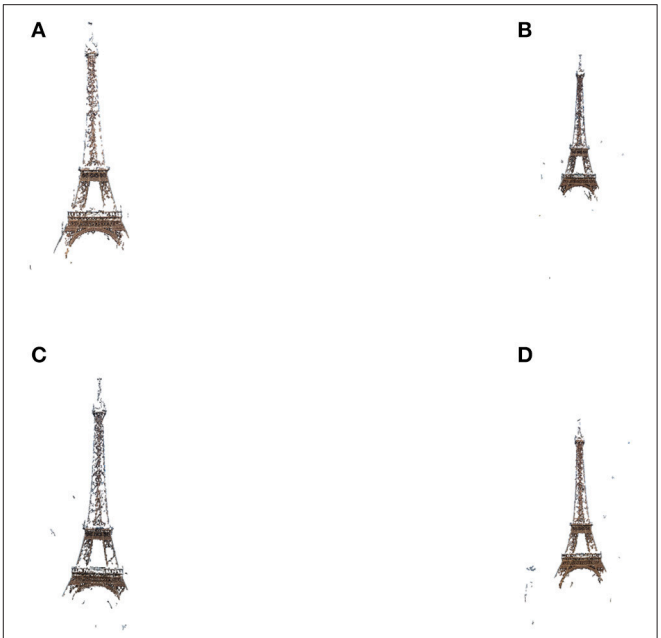

FIGURE 3 | The effect of the number of the outliers in reconstruction performance using the SfM algorithm. **(A)** 100 inliers, 0 outliers. **(B)** 90 inliers, 10 outliers. **(C)** 70 inliers, 30 outliers. **(D)** 60 inliers, 40 outliers.

TABLE 2 | Execution time for 3D reconstruction with respect to the number of fed images.

100 images	1013.2 s
90 images	883 s
80 images	715 s
70 images	635 s
60 images	571 s

As a result, in our approach about 100 images are considered adequate to provide a satisfied reconstruction of the monument. However, these images have been selected by removing the plethora our outliers and keeping only the most representative data as being automatically extracted by our algorithms.

Table 2 shows the time execution for the 3D reconstruction methods with respect to the number of images fed as inputs to the SfM. It is clear that as the number of images increases the respective required time also increases but also the reconstruction accuracy is improved as well. Thus, there is a trade-off between the requested time and 3D accuracy. This reveals the significance of our scheme. The target is to select a small number of frames from the short duration video shots that will represent the different orientation views of the monument as much as possible. Increasing the number of frames will lead to a greater cost while what we can achieve in precision of the reconstruction is more and more saturated though the increase in the number of frames used. Thus, if we need to reach a high detailed 3D reconstruction, we should take into account more frames representing different object views. On the contrary, if there is a time limitation, due for example to our device capabilities (e.g., mobile devices), a smaller but representative number of frames should be exploited to accelerate the process while keeping reconstruction precision as high as possible. We should stress that SfM is a polynomial complex algorithm and thus increasing the number of frames used as inputs the time is exponentially increased.

The Impact of the Proposed Scheme in Cultural Heritage

The overwhelming majority of tangible cultural heritage assets are located in regions where complete protection is not possible due to financial, environmental, political, religious or other local factors. Most of globe country are poor, pursuing the increase of the income and quality of life of their citizens and thus leaving protection of cultural heritage as a second option. In addition, regional poverty often goes with environmental decay of the soil, water and air. These pollutants will have a tremendous impact of object materials decaying their structures and thus putting cultural heritage in danger. On the other hand, local conflicts, wars, lootings, and other disputes frequently lead to partial or fully destruction of cultural objects with a great impact of culture and the local civilization.

Archaeologists, cultural heritage scientists and engineers need 3D geometric models of cultural heritage objects so as to derive documentation of them. However, funding 3D capturing procedures for all the plethora of cultural heritage monuments is not possible especially for poor or unstable countries. This gap is covered by the proposed scheme which exploits simple video shots mainly captured from touristic purposes or simple visits to derive 3D geometric models of the objects. The recent advances in hardware and software technology make video capturing devices be of low-cost and thus simple video shots be available for everyone, anytime and everywhere. Thus, a massive 3D documentation and protection can be achieved.

On the other hand, the derived 3D models can be useful for augmented reality (AR) applications triggering a new series of applications, such as games for promoting cultural heritage sites, overlay of natural with virtual objects for more precise documentation and relation of some cultural assets with others.

CONCLUSIONS

The today's dramatic decrease in the cost of capturing multimedia data has stimulated a great expansion of multimedia data which are stored, and processed over distributed and heterogeneous repositories. This results in a tremendous number of multimedia data which can be exploited to trigger several applications and launch new multimedia networked services.

One of the key advantages of this tremendous volume of multimedia information is to be exploited to 3D reconstruct objects of interest, monuments, site or other regions without the extra cost of processing or capturing the rich media content within high degrees of accuracy. Across most of the aforementioned multimedia repositories, the existence of short videos, mainly being captured for personal use, is a significant part of multimedia information which can be exploited for 3D processing. To identify the key frames, initially videos are clustered together with respect to their textual descriptions as derived from the caption annotation. Then, an outlier removal algorithm is proposed to make the pool of videos more homogeneous. The core part of the proposed scheme is the implementation of a novel video summarization scheme based on a discriminant Principal Component Analysis (d-PCA).

The experiments conducted on a large dataset of cultural objects indicate that the proposed algorithm (a) can 3D reconstruct sites or objects of interest even though the data have been obtained from unstructured visual content, (b) the proposed summarization scheme can accurately localize the data of interest than other approaches. The results indicate that even a small number of frames is adequate to reconstruct the objects of interest.

In future, we intend to expand this work in embedding time component in the reconstruction phase; that is, how a monument is evolved in time and in season. This will lead to a 4D reconstruction f (3D geometry plus time) implemented under a massive way (Kyriakaki et al., 2014). This will trigger a series of new applications both for the cultural experts

or for the simple users. For instance, the latter can share unique 3D experiences on how a monument is changing through different seasons under snow, rain, or hot conditions. The first can share some small geometric changes in the monuments that can assist them in documentation and analysis. Furthermore, the massive 3D reconstruction can boost a series of applications in Augmented Reality (AR) and Virtual Reality (VR) domain by superposing story telling algorithm with unique 3D objects.

4D modeling can be very useful for covering the intangible cultural heritage era and especially the digitalization of dances. A dance can be seen as a dynamic time evolved model and thus 4D reconstruction can be much more challenging and demanding (Aristidou et al., 2014, 2016). Specialized software toolkits needed to be applied for such digitization such as VICON (Rallis et al., 2017, 2018) while its unstructured modeling from UGC is really a very arduous task.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This work is supported by the European Union funded project H2020 TERPSICHORE Transforming Intangible Folkloric Performing Arts into Tangible Choreographic Digital Objects under grant agreement 691218.

REFERENCES

Alsadik, B., Gerke, M., Vosselman, G., Daham, A., and Jasim, L. (2014). Minimal camera networks for 3D image based modeling of cultural heritage objects. *Sensors* 14, 5785–5804. doi: 10.3390/s140405785

Ankerst, M., Breunig, M. M., Kriegel, H. P., and Sander, J. (1999). OPTICS: ordering points to identify the clustering structure. *ACM Sigmod Record* 28, 49–60, doi: 10.1145/304181.304187

Aristidou, A., Stavrakis, E., Charalambous, P., Chrysanthou, Y., and Himona, S. L. (2016). Folk dance evaluation using laban movement analysis. *J. Comput. Cult. Heritage* 8:20. doi: 10.1145/2755566

Aristidou, A., Stavrakis, E., and Chrysanthou, Y. (2014). Motion Analysis for Folk Dance Evaluation, in *GCH*, eds D. Fellner and R. Scopigno (Darmstadt), 55–64.

Avrithis, Y. S., Doulamis, A. D., Doulamis, N. D., and Kollias, S. D. (1999). A stochastic framework for optimal key frame extraction from MPEG video databases. *Comput. Vision Image Understand.* 75, 3–24, doi: 10.1006/cviu.1999.0761

Barazzetti, L., Scaioni, M., and Remondino, F. (2010). Orientation and 3D modeling from markerless terrestrial images: Combining accuracy with automation. *Photogrammetr.Record* 25, 356–381. doi: 10.1111/j.1477-9730.2010.00599.x

Bolles, R. C., Baker, H. H., and Marimont, D. H. (1987). Epipolar-plane image analysis: an approach to determining structure from motion. *Int. J. Comput. Vis.* 1, 7–55. doi: 10.1007/BF00128525

Bruno, F., Bruno, S., De Sensi, G., Luchi, M. L., Mancuso, S., and Muzzupappa, M. (2010). From 3D reconstruction to virtual reality: a complete methodology for digital archaeological exhibition. *J. Cult. Heritage* 11, 42–49. doi: 10.1016/j.culher.2009.02.006

Cernekova, Z., Pitas, I., and Nikou, C. (2006). Information theory-based shot cut/fade detection and video summarization. *IEEE Transact. Circ. Syst. Video Technol.* 16, 82–91. doi: 10.1109/TCSVT.2005.856896

Dorninger, P., and Nothegger, C. (2007). "3D segmentation of unstructured point clouds for building modeling," Proc. *of Photogrammetric Image Analysis (PIA)*, 191–196,

Doulamis, A., Doulamis, N., Protopapadakis, E., Voulodimos, A., and Ioannides, M. (2018). "4D modeling in cultural heritage," in *Advances in Digital Cultural Heritage*. (Cham: Springer), 174–196.

Doulamis, A., Ioannides, M., Doulamis, N., Hadjiprocopis, A., Fritsch, D., Balet, O., et al. (2013). 4D reconstruction of the past. *Proc. SPIE* 8795:87950J. doi: 10.1117/12.2029010

Doulamis, A. D., Doulamis, N., and Kollas, S. (2000b). Non-sequential video content representation using temporal variation of feature vectors. *IEEE Transact. Consumer Electron.* 46, 758–768. doi: 10.1109/30.883444

Doulamis, A. D., Doulamis, N. D., and Kollias, S. D. (2000a). Fuzzy video content representation for video summarization and content-based retrieval. *Signal Process.* 80, 1049–1067. doi: 10.1016/S0165-1684(00)00019-0

Doulamis, N., Yiakoumettis, C., Miaoulis, G., and Protopapadakis, E. (2013). "A constraint inductive learning-spectral clustering methodology for personalized 3D navigation," in *International Symposium on Visual Computing* (Berlin, Heidelberg: Springer).

Doulamis, N. D., Doulamis, A. D., Avrithis, Y. S., Ntalianis, K. S., and Kollias, S. D. (2000). Efficient summarization of stereoscopic video sequences. *IEEE Transact. Circ. Syst. Video Technol.* 10, 501–517. doi: 10.1109/76.844996

Doulamis, N. D., Doulamis, A. D., Kokkinos, P., and Varvarigos, E. (2016). Event detection in twitter microblogging. *IEEE Trans. Cybernet.* 46, 2810–2824, doi: 10.1109/TCYB.2015.2489841

Fritsch, D., and Klein, M. (2018). 3D preservation of buildings–Reconstructing the past. *Multimedia Tools Appl.* 77, 9153–9170. doi: 10.1007/s11042-017-4654-5

Gargallo, P., and Sturm, P. (2005). Bayesian 3D modeling from images using multiple depth maps. *Proc. IEEE Comput. Soc. Confer. Comput. Vis. Pattern Recogn.* 2, 885–891. doi: 10.1109/CVPR.2005.84

Georgousis, S., Stentoumis, C., Doulamis, N., and Voulodimos, A. (2016). "A hybrid algorithm for dense stereo correspondences in challenging indoor scenes," in *IEEE International Conference on Imaging Systems and Techniques, 460-465*, Chania.

Guo, L., Chen, X., Liu, B., and Liu, T. (2014). 3D-object reconstruction based on fusion of depth images by Kinect sensor. *J. Appl. Optics* 35, 811–816.

Halkos, D., Doulamis, N., and Doulamis, A. (2009). A secure framework exploiting content guided and automated algorithms for real time video searching. *Multimedia Tools Appl.* 42, 343–375. doi: 10.1007/s11042-008-0234-z

Hirschmüller, H. (2008). Stereo processing by semiglobal matching and mutual information. *IEEE Transac. Pattern Anal. Mach. Intell.* 30, 328–341. doi: 10.1109/TPAMI.2007.1166

Huang, H., Cagniart, C., Boyer, E., and Ilic, S. (2016). A bayesian approach to multi-view 4D modeling. *Int. J. Comput. Vis.* 116, 115–135. doi: 10.1007/s11263-015-0832-y

Ioannides, M., Hadjiprocopis, A., Doulamis, N., Doulamis, A., and E., Protopapadakis, et. al. (2013). Online 4D reconstruction using multi-images. *ISPRS Ann. Photogr. Remote Sens. Saptial Inform. Sci.* 1, 169–174. doi: 10.5194/isprsannals-II-5-W1-169-2013

Ioannidis, C., Potsiou, C., Soile, S., Verykokou, S., Mourafetis, G., and Doulamis, N. (2016). "Technical aspects for the creation of a multi-dimensional land information system," in *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* (Paphos), 41.

Ioannidis, C., Psaltis, C., and Potsiou, C. (2009). Towards a strategy for control of suburban informal buildings through automatic change detection. *Comput. Environ. Urban Syst.* 33, 64–74. doi: 10.1016/j.compenvurbsys.2008.09.010

Ji, Z., Zhang, Y., Pang, Y., and Li, X. (2018).Hypergraph dominant set based multi-video summarization. *Signal Process.* 148, 114–123. doi: 10.1016/j.sigpro.2018.01.028

Kim, H., Yoon, I., Kim, T., and Paik, J. (2016). "Video summarization using feature dissimilarity," in *International Conference on Electronics, Information, and Communications, ICEIC* (Da Nang).

Kim, W. H., Kim, H., Park, J. H., and Jeong, S. Y. (2014). Time pattern locking scheme for secure multimedia contents in human-centric device. *Sci. World J.* 2014:796515. doi: 10.1155/2014/796515

Kim, Y. M., Theobalt, C., Diebel, J., Kosecka, J., Miscusik, B., and Thrun, S. (2009). "Multi-view image and ToF sensor fusion for dense 3D reconstruction," in *IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops* (Kyoto), 1542–1546.

Kosmopoulos, D. I., Doulamis, A., Makris, A., Doulamis, N., Chatzis, S., and Middleton, S. E. (2009). Vision-based production of personalized video. *Signal Process. Image Commun.* 24, 158–176. doi: 10.1016/j.image.2008.12.010

Kuanar, S. K., Ranga, K. B., and Chowdhury, A. S. (2015). Multi-view video summarization using bipartite matching constrained optimum-path forest clustering. *IEEE Transact. Multimedia* 17, 1166–1173. doi: 10.1109/TMM.2015.2443558

Kyriakaki, G., Doulamis, A., Doulamis, N., Ioannides, M., Makantasis, K., Protopapadakis, E., et al. (2014). 4D reconstruction of tangible cultural heritage objects from web-retrieved images. *Int. J. Heritage Dig. Era* 3, 431–451. doi: 10.1260/2047-4970.3.2.431

Laggis, A., Doulamis, N., Protopapadakis, E., and Georgopoulos, A. (2017). A low-cost markerless tracking system for trajectory interpretation. *Int. Arch. Photogrammetry Remote Sens. Spatial Inform. Sci.* 42, 413–418. doi: 10.5194/isprs-archives-XLII-2-W3-413-2017

Li, Y., Zhang, Z., Peng, Y., Yin, H., and Xu, Q. (2018). Matching user accounts based on user generated content across social networks. *Future Generat. Comput. Syst.* 83, 104–115. doi: 10.1016/j.future.2018.01.041

Li, Z., Schuster, G. M., and Katsaggelos, A. K. (2005). MINMAX optimal video summarization. *IEEE Transact. Circ. Syst. Video Technol.* 15, 1245–1256. doi: 10.1109/TCSVT.2005.854230

Mademlis, I., Tefas, A., Nikolaidis, N., and Pitas, I. (2016). Multimodal stereoscopic movie summarization conforming to narrative characteristics. *IEEE Transact. Image Process.* 25, 5828–5840. doi: 10.1109/T.I.P.2016.2615289

Makantasis, K., Doulamis, A., Doulamis, N., and Ioannides, M. (2016). In the wild image retrieval and clustering for 3D cultural heritage landmarks reconstruction. *Multimedia Tools Appl.* 75, 3593–3629. doi: 10.1007/s11042-014-2191-z

Meng, J., Wang, S., Wang, H., Yuan, J., and Tan, Y. P. (2018). Video summarization via multiview representative selection. *IEEE Transact. Image Process.* 27, 2134–2145. doi: 10.1109/TIP.2017.2789332

Money, A. G., and Agius, H. (2008). Video summarisation: a conceptual framework and survey of the state of the art. *J. Visual Commun. Image Representation* 19, 121–143. doi: 10.1016/j.jvcir.2007.04.002

Mundur, P., Rao, Y., and Yesha, Y. (2006). Keyframe-based video summarization using Delaunay clustering. *Int. J. Dig. Libraries* 6, 219–232. doi: 10.1007/s00799-005-0129-9

Ngo, W. C., Ma, F. Y., and Zhang, J. H (2005). Video summarization and scene detection by graph modeling. *IEEE Transact. Circ. Syst. Video Technol.* 15, 296–304. doi: 10.1109/TCSVT.2004.841694

Nguyen, C. V., Izadi, S., and Lovell, D. (2012). "Modeling kinect sensor noise for improved 3D reconstruction and tracking," in *Proceedings - 2nd Joint 3DIM/3DPVT Conference: 3D Imaging, Modeling, Processing, Visualization and Transmission, 3DIMPVT 2012* (Zurich), 524–530.

Ntalianis, K., and Doulamis, N. (2016). An automatic event-complementing human life summarization scheme based on a social computing method over social media content. *Multimedia Tools Appl.* 75, 15123–15149. doi: 10.1007/s11042-015-2454-3

Panagiotakis, C., Doulamis, A., and Tziritas, G. (2009). Equivalent key frames selection based on iso-content principles. *IEEE Transact. Circ. Syst. Video Technol.* 19, 447–451. doi: 10.1109/TCSVT.2009.2013517

Panagiotakis, C., Grinias, I., and Tziritas, G. (2007). "MINMAX video summarization under equality principle," in *IEEE 9th International Workshop on Multimedia Signal Processing, MMSP* (Chania), 272–275.

Rallis, I., Doulamis, N., Doulamis, A., Voulodimos, A., and Vescoukis, V. (2018). Spatio-temporal summarization of dance choreographies. *Comput. Graph.* 73, 88–101. doi: 10.1016/j.cag.2018.04.003

Rallis, I., Georgoulas, I., Doulamis, N., Voulodimos, A., and Terzopoulos, P. (2017). "Extraction of key postures from 3D human motion data for choreography summarization," in *9th IEEE International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)* (Athens), 94–101.

Remondino, F., and Clive, F. (2005). "Digital camera calibration methods: considerations and comparisons," in International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. 36, ed International Society of Photogrammetry and Remote Sensing (Heipke: International Society of Photogrammetry and Remote Sensing), 266–272.

Remondino, F., and El-Hakim, S. (2006). Image-based 3D modeling: a review. *Photogrammetric Rec.* 21, 269–291. doi: 10.1111/j.1477-9730.2006.00383.x

Remondino, F., and Stylianidis, E. (2016). *3D Recording, Documentation and Management of Cultural Heritage*. Whittles Publishing.

Rothganger, F., Lazebnik, S., Schmid, C., and Ponce, J. (2006). 3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *Int. J. Comput. Vis.* 66, 231–259. doi: 10.1007/s11263-005-3674-1

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). "ORB: an efficient alternative to SIFT or SURF," in *IEEE International Conference on Computer Vision, ICCV* (Barcelona), 2564–2571.

Rutkowski, T. M., and Mandic, D. P. (2007). "Modeling the communication Atmosphere: a human centered multimedia approach to evaluate communicative situations," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4451 *LNAI*, 155–169.

Sanou, B. (2015). *"ICT Facts and Figures," ICT Data and Statistics Division, Telecommunication Development Bureau, International Telecommunication Union (ITU), Place des* Nations. Switzerland.

Schnabel, R., Wahl, R., and Klein, R. (2007). Efficient RANSAC for point-cloud shape detection. *Comput. Graph. Forum* 26, 214–226. doi: 10.1111/j.1467-8659.2007.01016.x

Sevillano, X., Piatrik, T., Chandramouli, K., Zhang, Q., and Izquierdo, E. (2012). Indexing large online multimedia repositories using semantic expansion and visual analysis. *IEEE Multimedia* 19, 53–61. doi: 10.1109/MMUL.2012.28

Smith, M., Szongott, C., Henne, B., and Von Voigt, G. (2012). "Big data privacy issues in public social media," in *6th IEEE International Conference on Digital Ecosystems Technologies (DEST)* (Campione d'Italia), 1-6.

Soursos, S., and Doulamis, N. (2012). "Connected TV and beyond," in *IEEE Consumer Communications and Networking Conference, CCNC* (Las Vegas, NV), 582–586.

Torresani, L., Hertzmann, A., and Bregler, C. (2008). Nonrigid structure-from-motion: estimating shape and motion with hierarchical priors. *IEEE Transact. Pattern Anal. Mach. Intell.* 30, 878–892. doi: 10.1109/TPAMI.2007.70752

Triggs, B., McLauchlan, P. F., Hartley, R. I., and Fitzgibbon, A. W. (1999). "Bundle adjustment—a modern synthesis," in *International Workshop on Vision Algorithms*. Berlin, Heidelberg: Springer. 298–372.

Verykokou, S., Ioannidis, C., Athanasiou, G., Doulamis, N., and Amditis, A. (2017). 3D reconstruction of disaster scenes for urban search and rescue. *Multimedia Tools Appl.* 77:9691. doi: 10.1007/s11042-017-5450-y

Vishnevskaya, E. V., Klimova, T. B., Bohomazov, I. V., Dumacheva, E. V., and Yakovenko, O. V. (2015). The importance of multimedia and interactive content for increasing tourist attractiveness of the territory. *Medit. J. Soc. Sci.* 6, 561–567. doi: 10.5901/mjss.2015.v6n4s1p561

Wang, G., Chen, J., and Giannakis, G., B. (2018). "DPCA: dimensionality reduction for discriminative analytics of multiple large-scale datasets,"in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary*.

Wang, S., and Dey, S. (2013). Adaptive mobile cloud computing to enable rich mobile multimedia applications. *IEEE Trans Multimedia* 15, 870–883. doi: 10.1109/TMM.2013.2240674

Westoby, M. J., Brasington, J., Glasser, N. F., Hambrey, M. J., and Reynolds, J. M. (2012). Structure-from-Motion' photogrammetry: a low-cost, effective tool for geoscience applications. *Geomorphology* 179, 300–314. doi: 10.1016/j.geomorph.2012.08.021

Xia, D., Yang, F., and Li, Q. (2013). Fast 3D modeling from images. *Optik* 124, 4621–4626. doi: 10.1016/j.ijleo.2013.01.090

Yastikli, N. (2007). Documentation of cultural heritage using digital photogrammetry and laser scanning. *J. Cult. Heritage* 8, 423–427. doi: 10.1016/j.culher.2007.06.003

Yiakoumettis, C., Doulamis, N., Miaoulis, G., and Ghazanfarpour, D. (2014). Active learning of user's preferences estimation towards a personalized 3D navigation of geo-referenced scenes. *GeoInformatica* 18, 27–62. doi: 10.1007/s10707-013-0176-0

Zhang, L., and Chen, X. (2014). "Topology-based automatic 3D modeling from multiple images," in *6th International Conference on Wireless Communications and Signal Processing, WCSP* (Hefei).

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.