Automatic Aerial Image Alignment for GeoMemories

Giuseppe Amato, Fabrizio Falchi, Fausto Rabitti
Information Science and Technology Institute (ISTI)
National Research Council (CNR)
Pisa, Italy
name.surname@isti.cnr.it

Andrea Marchetti, Maurizio Tesconi
Institute of Informatics and Telematics (IIT)
National Research Council (CNR)
Pisa, Italy
name.surname@iit.cnr.it

***Abstract*—In the last few years, aerial and satellite photographs have become more an more important for historical records. The availability of Geographical Information Systems and the increasing number of photos made per year allows very advanced fruition of large number of contents. In this paper we illustrate the GeoMemories approach and we focus on its automatic image alignment architecture. The approach leverages on a set of georeferenced images used as knowledge base. Local features are used in combination with compact codes and space transformation to achieve high level of efficiency.**

***Keywords*—image alignment; BoW; VLAD; local features; aerial photos;**

I. Introduction

GeoMemories[1] is a joint project between the AeroFototeca Nazionale (AFN) and the Institute of Informatics and Telematics (IIT) of the CNR. AFN has an extensive set of aerial photographs that constitute a historical record of the Italian territory from the end of the XIX century up to the end of the XX century. The project has created a web platform that allows browsing aerial photos traveling along the spatial-temporal dimensions. For efficient and effective managing such a huge archive, automatic alignment algorithms are crucial for placing historical aerial photos on top of nowadays maps.

During the last decade, local features have emerged, as an effective approach for image alignment, copy detection, landmark recognition, etc. A drawback of the use of local features is that a single image is represented by a possible large set of descriptors that should be individually matched and processed in order to compare the visual content of two images. In order to improve the efficiency of image matching on a large scale, a very popular method is the Bag of Features or Bag of Words (BoW) approach. This approach describes each image as a subset of predefined (visual) words. Thus, techniques successfully applied to text retrieval, can be easily applied also in the context of content based image retrieval.

Recently, another promising direction has emerged to simplify the representation of local features, based on the use of compact codes. The Vector of Locally Aggregated Descriptors (VLAD) [1] and its probabilistic counterpart (i.e., Fisher vectors) [2] compactly represent the local features of an image as a single fixed size vector. Image pairs can be then compared by using similarity (or distance) functions applied to the compact vectors. In this way, all the techniques extensively studied for efficient similarity search can be applied. In fact, an image query corresponds to a single fixed size vector [3]. For instance, in [4] it was shown that Euclidean Locality-Sensitive Hashing (LSH) [5] techniques can be efficiently and effectively applied with VLAD.

In this paper we propose a comprehensive approach for efficient alignment of aerial images available in the GeoMemories project. The approach consists of 3 stages with increasingly cost of the analysis but with decreasing number of candidate images. While the whole dataset is searched for finding similar images to the one to be georeferenced, in the subsequent steps more robust approach are considered only between the candidate set selected at the step before. In this way we are able to merge the high efficiency of the recognition algorithms developed in the Multimedia Information Retrieval field (i.e., VLAD and BoW) with the high effectiveness of Computer Vision approaches relying on local features and Random Sample Consensus (RANSAC) [6].

II. Background

A. Local Features

A local feature is an image pattern, which differs from its immediate neighborhood. It is usually associated with a change of an image property or several properties simultaneously, although it is not necessarily localized exactly on this change. Local features describe interesting regions in an image. Interesting regions differ from their immediate neighborhood and should be consistently identified on any two images representing the same visual content [7]. The description of each interesting region [8] have to be robust to region transformation such as scale, rotation, affine, homogrphy, etc...

The Scale Invariant Feature Transform (SIFT) [9], the most famous local feature, is a representation of the low level image content that is based on a transformation of the image data into scale-invariant coordinates relative to local features extracted from keypoints in an image. Keypoints are invariant to scale and orientation, selected by choosing the most stable points from a set of candidate location. Each keypoint in an image is associated with one or more orientations, based on local image gradients. Image matching is performed by comparing, typically using the Euclidean distance, the descriptions. Even if many other local features have been proposed in the last few year (e.g., SURF [10]) SIFT is still the most widely adopted.

[1] http://www.geomemories.org

B. Bag of Words (BoW)

The goal of the BoW approach [11] is to substitute each description of the region around an interest point (i.e., each local feature) of the images with visual words obtained from a predefined vocabulary in order to apply traditional text retrieval techniques to content-based image retrieval. The visual vocabulary is typically built selecting the centroids of clusters identified using as *k-means*. The second step is to assign each local feature of the image to the identifier of the first nearest word in the vocabulary. At the end of the process, each image is described as a set of visual words. The retrieval phase is then performed using text retrieval techniques considering a query image as a disjunctive text-query. Typically, the *cosine* similarity measure in conjunction with a term weighting scheme is adopted for evaluating the similarity between any two images.

In this paper we will not use BoW for indexing images but for finding candidate matching pairs between two images. In fact, for indexing, we will make use of the more efficiency and effective VLAD approach described below. Given a set of candidate matches, for effective image alignment it necessary to perform RANSAC [6] on candidate matching pairs of points between the query image and the referenced ones. Two speed up this process we make use of the BoW approach considering any region of interest described with the same word as matching.

C. Compact codes

Fisher kernels [12] describe how the set of descriptors deviates from an average distribution modeled by a parametric generative model. Fisher kernels have been applied in the context of image classification [13] and large scale image search [2]. In [4] it has been proved that Fisher vectors (FVs) extend the BoW. While the BoW approach counts the number of descriptors assigned to each region in the space, FV encodes the proximate location of the descriptors in each region and has a normalization that can be interpreted as an IDF term [14]. The FV image representation proposed by [13] assumes that the samples are distributed according to a Gaussian Mixture Model (GMM) estimated on a training set.

The VLAD representation was proposed in [1]. As for the BoW, a codebook $\{\mu_1, \dots, \mu_K\}$ is first learned using a cluster algorithm (e.g. k-means). Each local descriptor x_t is then associated to its nearest visual word $NN(x_t)$ in the codebook. For each codeword the differences between the vectors x_t assigned to μ_i are accumulated:

$$v_i = \sum_{x_t : NN(x_t) = i} x_t - \mu_i \qquad (1)$$

The VLAD is the concatenation of the accumulated vectors, i.e. $V = [v_1^T \dots v_K^T]$.

In order to compare two VLAD image representation with the inner product similarity function two normalization are performed: first, a power normalization with power 0.5; second, a L2 normalization. The observation that VLAD descriptions are relatively spare and very structured suggests a principal component analysis (PCA) that is usually performed to reduce the size of the Kd-dimensional VLAD vectors.

In [4], it has been proved that VLAD is a simplified non-probabilistic version of FV: VLAD is to FV what k-means is to GMM clustering. The k-means clustering can be viewed as a non-probabilistic limit case of GMM clustering. In [4] Euclidean Locality-Sensitive Hashing and its variant have been proposed to efficiently search VLAD descriptions.

D. Image alignment

Image alignment is the task of discovering the correspondence relationships among images with varying degrees of overlap. The survey given in [15] groups the approaches in the following categories: motion models [16], direct pixel-to-pixel comparisons, features based. In this work, we make use of a feature based approach. Our proposal is inspired by state-of-the art image registration (i.e., the process of transforming different sets of data into one coordinate system) in combination with high-efficiency techniques developed mainly for landmark recognition.

III. GEOMEMORIES

GeoMemories is a joint project between the AeroFototeca Nazionale (AFN) of the Italian Ministry of Cultural Heritage in Rome and the Institute of Informatics and Telematics (IIT) of the Consiglino Nazionale delle Ricerche (CNR). The project is funded by the Italian Internet Domains Registry. AFN has an extensive set of aerial photographs that constitute a large archive of memory creating a historical record of the Italian territory between XIX and XX centuries. This huge archive of some millions of photos consists of several collections among which the Royal Air Force and the USA Air Force ones represent a view of Italy as it was 70 years ago. This landscape no longer exists and it has been transformed by the post-war reconstruction, the economic boom, the modernization and some natural disasters.

The project has developed a web platform for browsing aerial photos traveling along the spatial-temporal dimensions with the opportunity to also integrate multimedia data from other open archives or from social contributions. Since Version 5.0 Google Earth has added a timeline to display historical imagery but this new feature is very basic and has few functionality. For instance, it is not possible to create fading effects between two historical maps of the same area. In addition, aerial photos provided by Google are relatively recent and the biggest limitation is the lack of relevant historical aerial photos. Italy, for example, has a significant coverage only from 2003, moreover the only samples of actual historical imagery (1943) concerns some cities (Rome, Florence, Naples, Turin, Trieste and Venice) and the image resolution is very low. Our aim is to rebuild a virtual globe as similar to Google Earth oriented to the management of the time providing historical and spatial information.

The aerial photos made available to the project are digitized and stored to form a parallel virtual archive: this is an important measure for the protection of the originals, all on paper or film, which over time can thus be withdrawn from the direct manipulation, and preserved in the best conservative way. Digital images are then subjected to different steps, as described by Fig. 2, to create historical maps. Each photo, after being digitized, cropped and eventually equalized to

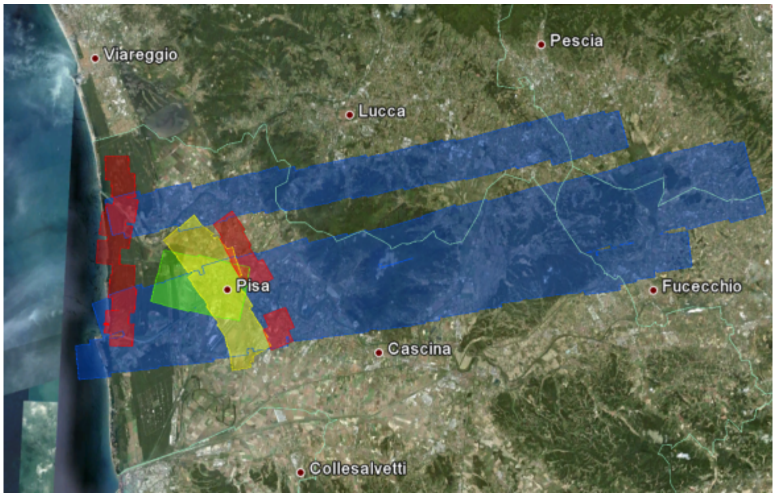

Fig. 1. Some aerial strip photos corresponding to 4 historical maps (20-08-1043 blue map, 18-02-1944 green map, 13-04-1944 yellow map and 14-05-1944 red map)

eliminate some exposure differences, are orthorectified and georeferenced. Google Imagery is used as reference map. Finally the georeferenced photos are joined by using mosaicing techniques.

Initially we have started to process 200 photos covering the northern part of Tuscany, and the years from 1943 to 1945, then from the second part of 2011 the set has been increased to 1000/2000 photos. This small set has been useful to develop and test several procedures to process the aerial strip photographs. At the end of the process we have realized four historical maps of the province of Pisa covering a time range from 20 August 1943 to 14 May 1944.

It is worth noting that the work flow for creating historical maps is really onerous and even though we can reduce the human component by developing automatic procedures, the huge size of the archive requires to find solutions based on social contributions so an important future activity of the project will concern the development of collaborative web applications to realize some steps such as georeferencing exploiting web volunteers (see the online tool http://www.georeferencer.org/).

The historical maps, each referring to a specific historical period, are browsable in the 4 spatio-temporal dimensions by using a web application based on Google Earth plugin and some javascript libraries. Fig. 3 shows a screenshot of the application. The user filter region and time of interest can select the corresponding historical maps and play with a fade in/out slide bar to display the evolution of the current area.

Google Earth integrates on its map different layers such as video, pictures, web cam by using the geographic reference (geotagging), we would like to use the same mechanism adding the time value (timetagging). The geo-historical data layers will come from the web obtained through web mining techniques, or filtered from open archives as wikipedia, youtube, flicker, and finally by raising social contributions related to initiatives for preserving historic memories. The result will be a sort of Historical Geographical Atlas where it will be possible to build tour to travel in the space-time dimension to tell stories.

A. Automatic Image Alignment

The image alignment task in GeoMemories consists in identifying overlapping of a non-georeferenced images with one of the georeferenced images already inserted in Geomemories. Our approach is features based and consists in retrieving first the most similar images in the knowledge based that are considered the best candidate for having an overlap with the query image.

As reported in [17] many geometric transformation can be found using the RANSAC algorthm [6]. Given the peculiar characteristics of our scenario, where the image dataset consists of aerial photos taken mostly from the vertical, we used the rotational and scale transformation, which provided us with the right compromise between simplicity and precision of results. In fact, even if the aerial photos represent a ground which is not strictly rigid and flat, more general transformations, such as affine and homography transformation, typically results in more noisy results and more difficult computation.

Given a point in a georeferenced image and a point in a non georeferenced image, we search for a rotational and scale transformation able to map the referenced point on the query image. Once such a transformation is found we are able to georeference the query image.

Unfortunately, comparing each aerial photos with subregions of the reference maps or another aerial photos using local features matching and RANSAC not only does not scale, but it is unfeasible for the scale of the problem we had. A single comparison takes seconds and the whole process would take hours. In GeoMemories, to speed-up the process, BoW and VLAD approaches are adopted for finding a set of candidates. Each image is processed as follows:

1) the SIFT local features are extracted using state of the art open source software (we used OpenIMAJ but also OpenCV could be used). In principle any other local feature such as SURF or ORB could be used and are under investigation.
2) word assignment for BoW approach is achieved comparing each local features with the vocabulary defined

ISBN: 978-1-61208-265-3

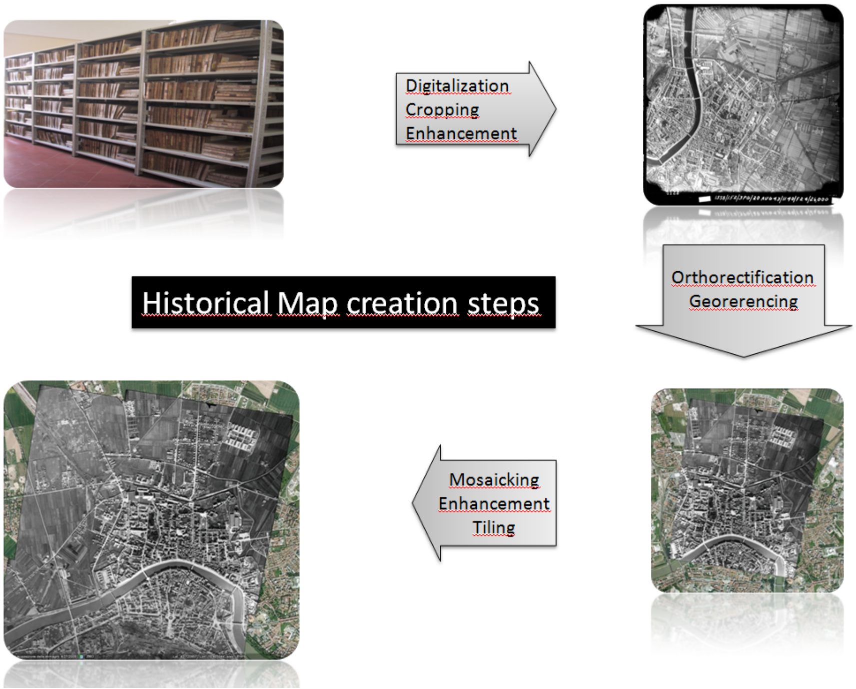

Fig. 2. The process steps for creating historical maps from aerial strip photography

using the k-means on a sample set of images. To speed-up the assignment kd-tree has been adopted.

3) the VLAD descriptions are obtained by comparing each local features with the small vocabulary (64 reference) that the VLAD requires. The VLAD description is then composed as described in Sec. II-C.

For efficiently finding candidate overlapping images, VLAD descriptions of each image was inserted in a similarity search index, namely the MI-File [18].

Given a not georeferenced image, the process of automatic alignment is performed as following:

1) local features, BoW and VLAD description are extracted as previously described
2) the VLAD description is used as a query for finding the 100 most similar georeferenced images in the index
3) the BoW description of the query image are compared with the 100 most similar images using RANSAC and searching for rotational and scale transformation
4) the 3 most promising images are then compared with the query using the SIFT local features and searching for a rotational and scale transformation using RANSAC

After each step the candidate set becomes smaller and more costly algorithms are used. While the second step consider all the images in the dataset but only relies in the VLAD descriptions without geometric consistency checks, at the third step RANSAC is used to check for overlaps. However at this stage local features are considered only for their label (the visual word assigned). Eventually, in the last step, the actual SIFTs are used for effective matching.

As mentioned before, both VLAD and BoW approaches require a set of reference local features to be selected between a knowledge base. In our experiments we considered a first set of 1000 aerial images and selected 128 local features for VLAD and 10.000 words to be used for BoW.

IV. Conclusion

In this work, we presented the GeoMemories project focusing on the automatic image alignment approach developed for georeferencing aerial photographs given a set of manually aligned images. The approach rely on a increasing knowledge based given that both automatic and manually georeferenced images inserted in GeoMemories are used for georeferencing subsequent images. Our approach efficiently and effectively aligns aerial photographs combing techniques from the Multimedia Information Retrieval field based on VLAD and BoW with high effective Computer Vision approach relying on local features and RANSAC. The approach consists of 3 stages with increasingly cost of the analysis but with decreasing number of candidate images. While the whole dataset is searched at first, in the subsequent steps more robust approach are considered only for the candidate set selected at the steps before. In this way we are able to merge the high efficiency of recognition approaches from the Multimedia Information Retrieval community with the high effectiveness of the algorithms developed in the Computer Vision field..

The proposed approach is actually under experimentation on the GeoMemories infrastructure. The low percentage of manually aligned photos available at this time, did not allow us to report experimental results in this preliminary work. However, subjective results are promising and we plan to report objective results in future works.

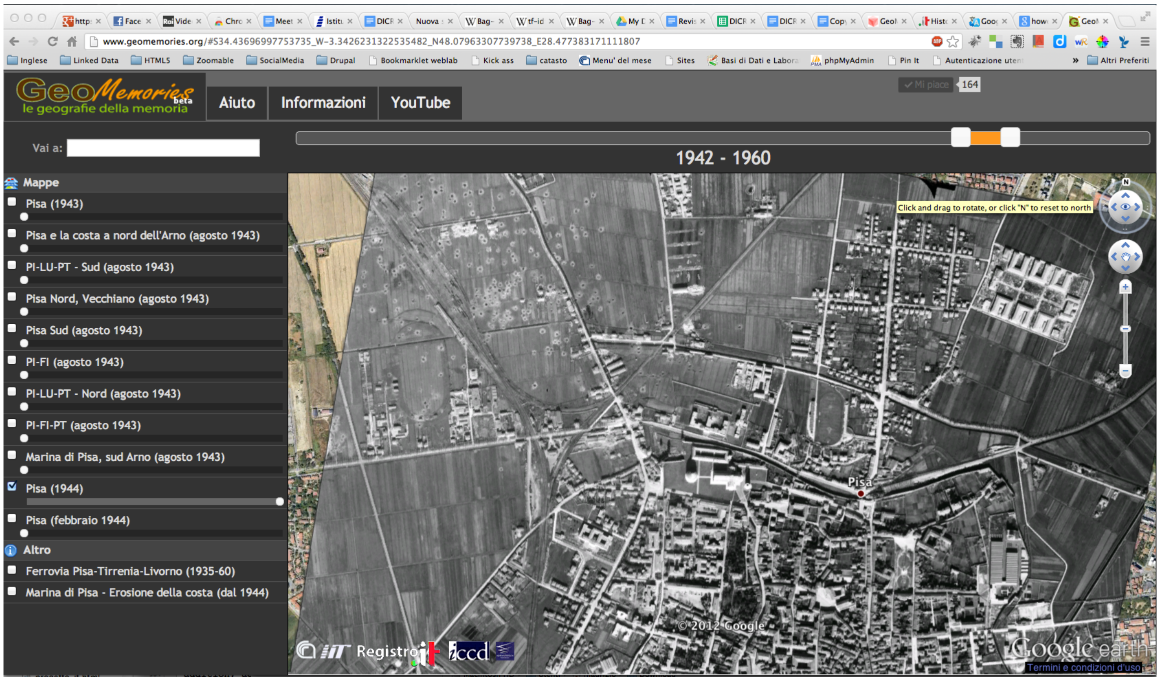
www.geomemories.org/#S34.43696997753735_W-3.3426231322535482_N48.07963307739738_E28.477383171111807
Inglese
Linked Data
HTML5
Zoomable
SocialMedia
Drupal
Bookmarklet weblab
Kick ass
catasto
Menu' del mese
Sites
Basi di Dati e Labora
phpMyAdmin
Pin It
Autenticazione uten
Altri Preferiti
GeoMemories beta
le geografie della memoria
Aiuto
Informazioni
YouTube
Mi piace
164
Vai a:
1942 - 1960
Mappe
Pisa (1943)
Pisa e la costa a nord dell'Arno (agosto 1943)
PI-LU-PT - Sud (agosto 1943)
Pisa Nord, Vecchiano (agosto 1943)
Pisa Sud (agosto 1943)
PI-FI (agosto 1943)
PI-LU-PT - Nord (agosto 1943)
PI-FI-PT (agosto 1943)
Marina di Pisa, sud Arno (agosto 1943)
Pisa (1944)
Pisa (febbraio 1944)
Altro
Ferrovia Pisa-Tirrenia-Livorno (1935-60)
Marina di Pisa - Erosione della costa (dal 1944)
Click and drag to rotate, or click "N" to reset to north
N
Pisa
© 2012 Google
Registro
iccd
Google earth
Termini e condizioni d'uso

Fig. 3. A screenshot of the first prototype for browsing the historical maps

References

[1] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *Int. J. Comput. Vision*, vol. 87, pp. 316–336, May 2010. [Online]. Available: http://dx.doi.org/10.1007/s11263-009-0285-2

[2] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, june 2010, pp. 3384 –3391.

[3] P. Zezula, G. Amato, V. Dohnal, and M. Batko, *Similarity Search: The Metric Space Approach*, ser. Advances in Database Systems. Springer-Verlag, 2006, vol. 32.

[4] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Sep. 2012, qUAERO. [Online]. Available: http://hal.inria.fr/inria-00633013

[5] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proceedings of the twentieth annual symposium on Computational geometry*, ser. SCG '04. New York, NY, USA: ACM, 2004, pp. 253–262.

[6] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[7] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Found. Trends. Comput. Graph. Vis.*, vol. 3, no. 3, pp. 177–280, 2008.

[8] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, 2005.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *In ECCV*, 2006, pp. 404–417.

[11] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ser. ICCV '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 1470–.

[12] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *In Advances in Neural Information Processing Systems 11*. MIT Press, 1998, pp. 487–493.

[13] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, june 2007, pp. 1 –8.

[14] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986.

[15] R. Szeliski, "Image alignment and stitching: a tutorial," *Found. Trends. Comput. Graph. Vis.*, vol. 2, no. 1, pp. 1–104, Jan. 2006.

[16] r. Szeliski, "Video mosaics for virtual environments," *IEEE Comput. Graph. Appl.*, vol. 16, no. 2, pp. 22–30, Mar. 1996.

[17] G. Amato, F. Falchi, and C. Gennaro, "Geometric consistency checks for knn based image classification relying on local features," in *SISAP '11: Fourth International Conference on Similarity Search and Applications, SISAP 2011, Lipari Island, Italy, June 30 - July 01, 2011*. ACM, 2011, pp. 81–88.

[18] G. Amato and P. Savino, "Approximate similarity search in metric spaces using inverted files," in *Proceedings of the 3rd international conference on Scalable information systems*, ser. InfoScale '08. ICST, Brussels, Belgium, Belgium: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2008, pp. 28:1–28:10. [Online]. Available: http://dl.acm.org/citation.cfm?id=1459693.1459731