Shortened Persistent Homology for a Biomedical Retrieval System with Relevance Feedback

Alessia Angeli, Massimo Ferri, Eleonora Monti, Ivan Tomba

▶ **To cite this version:**

Alessia Angeli, Massimo Ferri, Eleonora Monti, Ivan Tomba. Shortened Persistent Homology for a Biomedical Retrieval System with Relevance Feedback. 2nd International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE), Aug 2018, Hamburg, Germany. pp.282-292, 10.1007/978-3-319-99740-7_20 . hal-02060046

HAL Id: hal-02060046

https://inria.hal.science/hal-02060046v1

Submitted on 7 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Shortened persistent homology for a biomedical retrieval system with relevance feedback *

Alessia Angeli[1], Massimo Ferri[2], Eleonora Monti[1], and Ivan Tomba[3]

[1] Dept. of Mathematics, Univ. Bologna, Italy
{alessia.angeli,eleonora.monti5}@studio.unibo.it
[2] Dept. of Mathematics and ARCES, Univ. Bologna, Italy massimo.ferri@unibo.it
[3] 2R&D Dept., CA-MI S.r.l., Via Ugo La Malfa 13, Pilastro di Langhirano PR, Italy
tomba.ivan@gmail.com

Abstract. This is the report of a preliminary study, in which a new coding of persistence diagrams and two relevance feedback methods, designed for use with persistent homology, are combined. The coding consists in substituting persistence diagrams with complex polynomials; these are "shortened", in the sense that only the first few coefficients are used. The relevance feedback methods play on the user's feedback for changing the impact of the different filtering functions in determining the output.

Keywords: Persistence diagram · Elementary symmetric function · Projected gradient.

1 Introduction

The interaction between a medical doctor and a smart machine must respect at least two requirements: Fast action and good integration with the human operator. As far as mere morphology is concerned, deep learning has already reached performances comparable with the ones of a dermatologist [6]. In real practice, the number of (hidden or evident, formal or intuitive) parameters in a diagnostic task is so high, that a good synthesis in short times is, at least for the moment, the field of a human expert; the machine can anyway offer a reliable, stable, powerful assistance. It is necessary that the medical doctor understands what's going on in his/her interaction with the machine: This is a primary issue in the design of "explainable" AI systems [15]. A way out of the "black box" frustration is to accept a feedback from the user, so that the system adapts more and more to his/her viewpoint.

This is the case of a system currently developed by Ca-Mi srl, an Italian company producing biomedical devices, in collaboration with the Universities of Bologna and Parma and with the Romagna Institute for Study and Cure of Tumors (IRST). The machine acquires the image of a dermatological lesion and retrieves a set of most similar images out of a database with sure diagnoses.

* Article written within the activity of INdAM-GNSAGA.

The similarity is assessed by a relatively recent geometric-topological technique: Persistent homology. This tool is very effective above all on data of natural origin [8], but the main tool for classification and (dis)similarity assessment—the bottleneck distance between persistence diagrams—is computationally heavy. We are then experimenting a different coding of the same information contained in a persistence diagram, through a complex polynomial; the coding is then "shortened" in that we just use the first few coefficients, so that comparison and search becomes much faster; it appears that these first coefficients contain most of the relevant information.

While the first experiments are very promising [10], there is still a wide gap between what the system and the doctor see as "similar". Therefore it is an active area of research, to study a relevance feedback method for drawing the machine's formalization of similarity near the doctor's skilled view.

In this paper we present a preliminary study on a small public database (PH^2) of nevi and melanomas; our goal is to combine two relevance feedback methods expressly designed for persistent homology, with a new coding of one of its main tools, persistence diagrams.

Content Based Image retrieval (CBIR) is a ripe and challenging research area [11]. Apart from annotation-based systems, much of the success of CBIR is tied with the use of histograms (of colours, directions, etc.). Topological descriptors have entered the game but are not yet fully employed [19]. On the other hand, we are aware of the incredible flourishing of Deep Learning in the areas of image understanding and of medical diagnosis; our interest in the techniques proposed here depends on the needs to work with a rather limited database, and to be able to control step-by-step how the system adapts to the user. Still we plan, as a future step, to combine persistent homology with Deep Learning—as several researchers already do—by feeding a neural net with persistence diagrams. There at least two reasons for wishing such a development: The importance of the user's viewpoint (or taste, or goal) formalized by the choice of filtering functions; and the fact that whatever the kind of input, persistence diagrams have always the same structure, so that a learning network trained on persistence diagrams becomes immediately a much more versatile tool. A step in this direction has already been done [14].

2 Persistent homology

Persistent homology is a branch of computational topology, of remarkable success in shape analysis and pattern recognition. Its key idea is to analyze data through *filtering functions*, i.e. continuous functions f defined on a suitable topological space X with values e.g. in $\mathbb{R}$ (but sometimes in $\mathbb{R}^n$ or in a circle). Given a pair (X, f), with $f : X \to \mathbb{R}$ continuous, for each $u \in \mathbb{R}$ the *sublevel set* X_u is the set of elements of X whose value through f is less than or equal to u.

For each X_u one can compute the *homology modules* $H_r(X_u)$, vector spaces which summarize the presence of "voids" of dimension r, with $r \in \mathbb{N}$ (connected components in the case $r = 0$) and their relations. As there exist various homol-

ogy theories, some additional hypotheses might be requested on f depending on the choice of the homology.

Of course, if $u < v$ then $X_u \subseteq X_v$. There corresponds a linear map $\iota^r_{(u,v)} : H_r(X_u) \to H_r(X_v)$. On $\Delta^+ = \{(u,v) \in \mathbb{R}^2 \,|\, u < v\}$ we can then define the *r-Persistent Betti Number* (r-PBN) function

$$\begin{aligned} \beta^r_{(X,f)} : \ \Delta^+ \ &\to \quad \mathbb{Z} \\ (u,v) &\mapsto \dim \mathrm{Im}(\iota^r_{(u,v)}) \end{aligned}$$

All information carried by r-PBN's is condensed in some points (dubbed *proper cornerpoints*) and some half-lines (*cornerlines*); cornerlines are actually thought of as *cornerpoints at infinity*. Cornerpoints (proper and at infinity) build what is called the *persistence diagram* relative to dimension r. Figure 1 shows a letter "M" as space X, ordinate as function f on the left, its 0-PBN function at the center and the corresponding persistence diagram on the right.

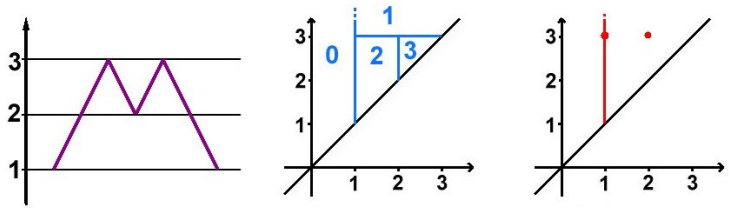

Fig. 1: Letter M, its 0-PBM function and the corresponding persistence diagram, relative to filtering function ordinate.

Remark 1. The theory also contemplates a *multiplicity* for cornerpoints (proper and at infinity); multiplicity higher than one is generally due to symmetries. We don't care about it in the present research, since all cornerpoints in our experiments have multiplicity one, as usual in diagrams coming from natural images.

Classification and retrieval of persistence diagrams (and consequently of the object they represent) is usually performed by the following distance, where persistence diagrams are completed by all points on the "diagonal" $\Delta = \{(u,v) \in \mathbb{R} \,|\, u = v\}$.

Definition 1. Bottleneck *(or* matching*)* distance.
Let $\mathcal{D}_k$ and $\mathcal{D'}_k$ be two persistence diagrams with a finite number of cornerpoints, the bottleneck distance $d_B(\mathcal{D}_k, \mathcal{D'}_k)$ *is defined as*

$$d_B(\mathcal{D}_k, \mathcal{D'}_k) = \min_{\sigma} \max_{P \in \mathcal{D}_k} \hat{d}(P, \sigma(P))$$

where σ varies among all the bijections between $\mathcal{D}_k$ and $\mathcal{D'}_k$ and

$$\hat{d}((u,v),(u',v')) = \min\left\{\max\left\{|u-u'|,|v-v'|\right\}, \max\left\{\frac{v-u}{2},\frac{v'-u'}{2}\right\}\right\}$$

given $(u,v) \in \mathcal{D}_k$ and $(u',v') \in \mathcal{D'}_k$.

For homology theory one can consult any text on algebraic topology, e.g. [13]. For persistent homology, two good references are [4, 5].

3 Symmetric functions of warped persistence diagrams

There are two main difficulties in comparing persistence diagrams. One is the fact that the diagonal Δ has a special role: For a persistence diagram, the diagonal Δ is a sort of "blown up" point, in the sense that points close to it are seen as close to each other by the bottleneck distance; moreover cornerpoints close to Δ generally represent noise, and it would be desirable to diminish their contribution. A second difficulty consists in the coding of a set of points in the plane, in a form suited for comparison and processing: E.g. an intuitive coding like making a $2M$ vector out of a set of M points is highly unstable and is not the solution to the problem of avoiding the permutations required by the bottleneck distance; a distance computed component by component might be very far from the otimal one. In [1]—which also contains a comparison with the bottleneck distance— we have tried to overcome both difficulties.

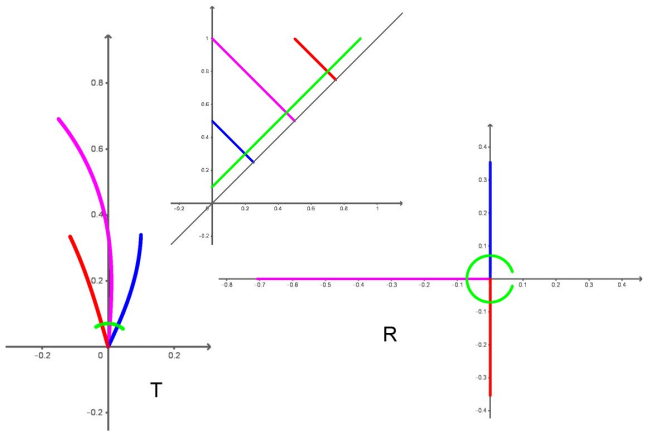

Fig. 2: The action of transformations T (left) and R (right) on some segments (above).

Following [9, 3], we have faced the first problem by two different transformations T (designed by Barbara Di Fabio) and R which "warp" the plane, so that all Δ is sent to $(0,0)$, seen here as the complex number zero:

$$T : \bar{\Delta}^+ \to \mathbb{C}, \quad T(u,v) = \tfrac{v-u}{2}(\cos(\alpha) - \sin(\alpha) + i(\cos(\alpha) + \sin(\alpha)))$$

where $\alpha = \sqrt{u^2 + v^2}$.

$$R : \bar{\Delta}^+ \to \mathbb{C}, \quad R(u,v) = \tfrac{v-u}{\sqrt{2}}(\cos(\theta) + i\sin(\theta))$$

where $\theta = \pi(u+v)$.

The main ideas behind T and R are: For the reasons mentioned at the beginning of this section, cornerpoints close to Δ ought to be considered close to each other. Therefore T and R (and other maps under study) take Δ to zero. Moreover they wrap the strip adjacent to Δ around zero itself, so that the contributions of noise cornerpoints should balance away in Viète's symmetric functions. We are still looking for the best wrapping function. Both T and R are continuous maps. See Figure 3 for a persistent diagram and its two images through T and R.

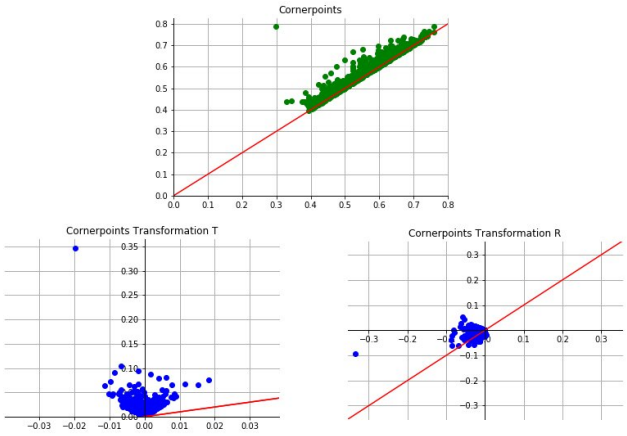

Fig. 3: A persistence diagram (above) and its images through T (left) and R.

As for the second problem, following an idea of Claudia Landi [9, 3] we decided to encode each (transformed) persistence diagram $\mathcal{D}$ as the polynomial having the complex numbers, images of the cornerpoints, as roots. We can then design distances built on the pairs of coefficients of equal position in the polynomials representing two diagrams, avoiding a combinatorial explosion. Actually, for a given diagram we form the vector having as components the elementary symmetric functions of the transformed cornerpoints (which, through Viète's formulas, equal the coefficients of the polynomial up to the sign) [16, Sect. IV.8]. So, the first component of the vector is the sum of all those numbers, the second one is the sum of all pairwise products, and so on. In order to take also cornerpoints at infinity into account, we have performed the following substitution.

Given a cornerpoint at infinity (i.e. a cornerline) with abscissa w of a persistence diagram $\mathcal{D}$, we substitute it with the point

$$\big(w, \max\{v \,|\, (u,v) \text{ is a cornerpoint of } \mathcal{D}\}\big)$$

Remark 2. Since cornerpoints near Δ generally represent noise, the two transformations were designed to wrap them around zero, so that the symmetric functions be scarcely affected by them. To this goal, transformation R fits better in our case, since our filtering functions (hence the cornerpoint coordinates) are bounded between 0 and 1.

For a fixed filtering function, for a fixed transformation (T or R), the same elementary symmetric function computed on two persistence diagrams may show a difference of several orders of magnitude, when there are many cornerpoints. This might consistently alter the comparisons, so we actually formed, for each persistence diagram $\mathcal{D}$, the complex vector $a_{\mathcal{D}}$ whose i-th component $a_{\mathcal{D}}(i)$ is the i-th root of the i-th elementary symmetric function of the transformed cornerpoints, divided by the number of cornerpoints. As an example, these are the (approximated) real parts of the first 10 symmetric functions for images IMD251 and IMD423, fitering function "Light intensity 2", transformation T:

IMD251:
(2E-1, -5.9E-1, 8.3E-2, 6.8E-2, -3.4E-2, 7.6E-3, -1E-3, 8.4E-5, -1E-6, -8.5E-7)
IMD423:
(7.1, -1.9E+2, -1.4E+3, 2.3E+3, 4.1E+4, 6E+4, -3.5E+5, -1.1E+6, 6.3E+5, 7.2E+6)

which, by taking the i-th root of the ith symmetric function, become:

IMD251:
(2E-1, -7.7E-1, 4.4E-1, 5.1E-1, -5.1E-1, 4.4E-1, -3.7E-1, 3.1E-1, -2.2E-1, -2.6E-1)
IMD423:
(7.1, -1.4E+1, -1.1E+1, 6.9, 8.4, 6.2, -6.2, -5.7, 4.4, 4.9)

and division by the number of cornerpoints (here 257 and 1408 respectively) yields:

IMD251:
(7.6E-4, -3E-3, 1.6E-3, 2E-3, -2E-3, 1.7E-3, -1.4E-3, 1.2E-3, -8.3E-4, -9.6E-4)
IMD423:
(5.1E-3, -9.7E-3, -8E-3, 4.9E-3, 5.9E-3, 4.4E-3, -4.4E-3, -4.1E-3, 3.1E-3, 3.4E-3)

whose differences finally make sense.

The advantage of using the vectors $a_{\mathcal{D}}$ instead of the original diagrams is that one can directly design a distance between complex vectors component-by-component (e.g. the L^1 distance we used), instead of considering all bijections between cornerpoint sets. Still, the computation of all the elementary symmetric functions would be too time consuming, so we performed in [1] some experiments by reducing the computation to the first k components, $k \in \{5, 10, 20, 50\}$, a small number compared with the hundreds of cornerpoints commonly found in the persistence diagrams of the examined images. Classification and retrieval of dermatological images using such "shortened" vectors was quite satisfactory with a dramatic time reduction.

4 Modifying distances

One of the major advantages of persistent homology is its modularity: By changing filtering function we change point of view on the shape of the objects under study and on the comparison criteria. Therefore the same data can be transformed into several pairs (X, f), and for for each pair we obtain a distance reflecting the features of X captured by filtering function f. Both for classification and for retrieval, we need a single distance, so we have to blend the distances we have into one. There are two rather natural choices for that: either the maximum or the arithmetic average.

Maximum is the initial choice of [12]. An ongoing research, by some of the authors of the present paper, prefers the average instead [17]; this agrees with the idea of cooperation of the different filtering functions (like in [2, 7]), versus one of them prevailing. We make the "neutral" choice of equal weights in the starting average for initializing the subsequent optimization process. As hinted in the Introduction, the research is aimed at enhancing a device of acquisition and retrieval of dermatological images. The concept of "similarity" is (and has to remain!) highly subjective in the medical domain: We want to adapt the system to the physician, not the other way around. This can be done by modifying the weights of the different distances when building a single distance, to approximate the (pseudo)distance δ representing the dissimilarity as perceived by the user.

Our setting is: We have a set $X = \{x_1, \ldots, x_N\}$ of objects (in our case dermatological images) and J descriptors $d^{(1)}, \ldots, d^{(J)}$ which give rise to an initial distance D^{IN} between the objects in X. In our study, D^{IN} is one of these two:

$$D^{MAX} = \max\{d^{(1)}, \ldots, d^{(J)}\}, \quad D^{AVG} = \frac{d^{(1)} + \cdots + d^{(J)}}{J}$$

Given a query q (i.e. an acquired image), the system retrieves the L objects closest to q with respect to D^{IN}, i.e. an L-tuple $X_q = (x_{i_1}, \ldots, x_{i_L})$ such that $D^{IN}(q, x_{i_1}) \leq \cdots \leq D^{IN}(q, x_{i_L})$. The user is shown these objects and expresses his/her relevance feedback by assessing the perceived dissimilarities as numbers $\delta(q, x_{i_1}), \ldots, \delta(q, x_{i_L})$.

While the Multilevel Relevance Feedback (MLRF) method proposed in [12] starts from D^{MAX}, then rescales the distances $d^{(i)}$ and takes the maximum, our Least Squares Relevance Feedback (LSRF) scheme start from D^{AVG} and computes a new distance D^{OUT} as

$$D^{OUT} = \sum_{j=1}^{J} \lambda_j d^{(j)}, \quad \lambda_j \geq 0$$

by minimizing the objective function

$$g(\lambda) = \|\mathbf{d}\lambda - \delta\|_2^2$$

i.e. by looking for $\lambda = \text{argmin}\|\mathbf{d}\lambda - \delta\|_2^2$, where the t-th row of matrix $\mathbf{d}$ is formed by the distances $d^{(j)}(q, x_{i_t})$, λ is the column matrix of the λ_j and δ is the column matrix formed by $\delta(q, x_{i_t})$ for $t = 1, \ldots, L$ and $j = 1, \ldots, J$.

Since this minimization problem might have multiple solutions, the vector of weights λ_j in D^{OUT} is obtained by iterating the Projected Gradient method. D^{OUT} is the best possible distance approximating the users similarity distance δ from the given data in a least-square sense. More details on this procedure will be given in an article to come.

5 Experimental results

As a preliminary study, we have tried to combine the modularity of persistent homology with the fast computation of the short vectors of symmetric functions of the transformed cornerpoints, with the adapting weights of relevance feedback.

We experimented with a small public database, PH^2 [18], containing 8-bit RGB, 768×560 pixels images of 80 common nevi, 80 atypical nevi, and 40 melanomas. We used 19 distances: 8 coming from simple morphological parameters, 11 coming from as many filtering functions; see Table 1 for a description of these features [10, 1]. It should be mentioned that the 11 distances coming from persistent homology already yield very good results, but the simple (and very fast computable) ones, obtained from morphological parameters, refine the general performance. For each image, we built the 11 persistence diagrams, performed one of the transformations T and R (see Sect. 3), and computed the 11 corresponding vectors, limited to length $k = 20$.

persistence features	**morphological features**
Light Intensity	Colour Histogram
Blue	Form Factor
Green	Haralick's Circularity
Red	Asymmetry
Excess Blue	Ellipticity
Excess Green	Eccentricity
Excess Red	Diameter
Light Intensity 2	Colour Entropy
Boundary Light Intensity	
Boundary	
Boundary 2	

Table 1: Features.

In normal operation, the user will give his/her feedback as follows. For $j = 1, \ldots, L$, the user is asked to assign a similarity judgement (where 0 stands for "dissimilar" and 1 for "similar") to the pair (q, x_{i_j}) with respect to three different

aspects of the skin lesion (boundary/shape, colours, texture). This results in a similarity vote v_{i_j} in the range $\{0,1,2,3\}$ which is transformed into $\delta(q, x_{i_j})$ by the following formula:

$$\delta(q, x_{i_j}) = \max\left\{0, \quad D^{IN}(q, x_{i_1}) + \frac{4(3-v_{i_j})-1}{10}\left(D^{IN}(q, x_{i_1}) - D^{IN}(q, x_{i_L})\right)\right\}$$

The formula assigns the values of δ in such a way that if the similarity vote v_{i_j} is maximal (3/3), then the corresponding δ is lower than the lowest value of the distances of the database images from the query and, conversely, if v_{i_j} is minimal, then δ is higher then the distance of the image which is farthest from it.

We finally retrieve again the L closest objects according to the modified distance (see Section 4).

In the present research $L = 10$ and the retrieval is performed by the "leave one out" scheme. Not having a real relevance feedback by a physician, we assign a retrieved image the maximal similarity vote $v = 3$ if it shares the same histological diagnosis of the query, and the minimal similarity vote $v = 0$ otherwise.

Assessment of the retrieval is performed by counting how many, of the retrieved lesions, have the same diagnosis of the query before and after the feedback. This is summarized in Table 2 by summing these scores for all images of the database as queries, starting from D^{MAX} and then applying the MLRF scheme, starting from D^{AVG} and applying our LSRF method.

As we can see, transformations R and T yield similar results. D^{AVG} performs better than D^{MAX} and LSRF produces a better improvement than MLRF.

	with D^{MAX}	**MLRF**	**difference**	**with D^{AVG}**	**LSRF**	**difference**
Transf. T	1694	1776	82	1729	1856	127
Transf. R	1869	1762	73	1735	1857	122

Table 2: Sums of scores with the two considered methods of relevance feedback.

6 Conclusions

We have compared—on a small public database of nevi and melanomas—two relevance feedback methods, conceived for a retrieval system based on persistent homology, combined with a computation reduction based on elementary symmetric functions of warped persistence diagrams. A weighted average of distances, with weights obtained by a standard optimization method, seems to perform better. The experiment will be extended to larger databases and with real feedback from expert dermatologists.

Acknowledgment

We wish to thank all the Reviewers for the very detailed and useful comments.

References

1. A. Angeli, M. Ferri, and I. Tomba. Symmetric functions for fast image retrieval with persistent homology. *preprint*, 2018.
2. A. Brucale, F. Cesari, M. d'Amico, M. Ferri, P. Frosini, L. Gualandri, M. Guerra, A. Lovato, and I. Pace. Image retrieval through abstract shape indication. In *Proceedings of the IAPR Workshop MVA 2000, Tokyo Nov. 28-30*, pages 367–370, 2000.
3. B. Di Fabio and M. Ferri. Comparing persistence diagrams through complex vectors. In *International Conference on Image Analysis and Processing*, pages 294–305. Springer, 2015.
4. H. Edelsbrunner and J. Harer. Persistent homology—a survey. In *Surveys on discrete and computational geometry*, volume 453 of *Contemp. Math.*, pages 257–282. Amer. Math. Soc., Providence, RI, 2008.
5. H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2009.
6. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.
7. M. Ferri. Graphic-based concept retrieval. In *International Conference on Availability, Reliability, and Security*, pages 460–468. Springer, 2013.
8. M. Ferri. *Persistent Topology for Natural Data Analysis — A Survey*, pages 117–133. Springer International Publishing, Cham, 2017.
9. M. Ferri and C. Landi. Representing size functions by complex polynomials. In *Proc. Math. Met. in Pattern Recognition*, volume 9, pages 16–19, 1999.
10. M. Ferri, I. Tomba, A. Visotti, and I. Stanganelli. A feasibility study for a persistent homology-based k-nearest neighbor search algorithm in melanoma detection. *Journal of Mathematical Imaging and Vision*, 57(3):324–339, 2017.
11. N. Ghosh, S. Agrawal, and M. Motwani. A survey of feature extraction for content-based image retrieval system. In *Proceedings of International Conference on Recent Advancement on Computer and Communication*, pages 305–313. Springer, 2018.
12. D. Giorgi, P. Frosini, M. Spagnuolo, and B. Falcidieno. 3D relevance feedback via multilevel relevance judgements. *The Visual Computer*, 26(10):1321–1338, 2010.
13. A. Hatcher. *Algebraic topology. 2002*, volume 606. 2002.
14. C. Hofer, R. Kwitt, M. Niethammer, and A. Uhl. Deep learning with topological signatures. In *Advances in Neural Information Processing Systems*, pages 1633–1643, 2017.
15. A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.
16. S. Lang. *Undergraduate algebra*. Springer Science & Business Media, 2005.
17. S. Magi, L. Mazzoni, E. Monti, I. Stanganelli, I. Tomba, and A. Visotti. Relevance feedback in a dermatology application. 2018. submitted.
18. T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira. Ph 2-a dermoscopic image database for research and benchmarking. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pages 5437–5440. IEEE, 2013.
19. M. Zeppelzauer, B. Zieliński, M. Juda, and M. Seidl. A study on topological descriptors for the analysis of 3d surface texture. *Computer Vision and Image Understanding*, 2017.