Large-Scale Endoscopic Image and Video Linking with Gradient-based Signatures

Christian Beecks
RWTH Aachen University, Germany
beecks@cs.rwth-aachen.de

Sabrina Kletz
Klagenfurt University, Austria
sabrina@itec.aau.at

Klaus Schoeffmann
Klagenfurt University, Austria
ks@itec.aau.at

***Abstract*—Given a large-scale video archive of surgical interventions and a medical image showing a specific moment of an operation, how to find the most image-related videos efficiently without the utilization of additional semantic characteristics? In this paper, we investigate a novel content-based approach of linking medical images with relevant video segments arising from endoscopic procedures. We propose to approximate the video segments' content-based features by *gradient-based signatures* and to index these signatures with the Minkowski distance in order to determine the most query-like video segments efficiently. We benchmark our approach on a large endoscopic image and video archive and show that our approach achieves a significant improvement in efficiency in comparison to the state-of-the-art while maintaining high accuracy.**

***Keywords*-medical image and video databases; endoscopy; content-based retrieval; gradient-based signatures**

I. Introduction

With the rapid spread of modern image and video cameras and their powerful processing capabilities as well as the advancement of digital image and video processing technology, more and more medical examinations and surgical interventions are recorded and stored in large-scale heterogeneous medical video archives. In fact, these medical archives have become increasingly popular in the field of medicine and healthcare in order to enrich the medical knowledge and decision process and to gain further information about and novel insight into the patients' surgical interventions. They are used, for instance, to provide patients information about their personal medical intervention or to train and teach young surgeons new operation techniques [1].

Though the digitization of medical examinations and surgical interventions provides a powerful information source for doctors and patients, the vendor and media diversity makes it difficult to unify the process of information integration and access. Moreover, the lack of semantic annotation – including the patients reference number or even the date of recording – inevitably leads to the question of how to efficiently access and inspect medical archives in a content-based way. Thus, given a medical image showing a specific moment of an operation, how to find the most similar video segments in a medical archive efficiently without the utilization of additional semantic characteristics?

In this paper we address the problem of accessing endoscopic video archives in a content-based way in order to efficiently link endoscopic images with the most relevant video segments. For this purpose, we approximate the video segments' content-based properties by means of *gradient-based signatures* [2] and index these signatures with the Minkowski distance. Gradient-based signatures epitomize a parameter-based content representation based on a joint generative model. In this way, our approach is customizable with respect to different types of media and content-based features and compatibly with existing solutions [1]. Moreover, this approach allows us to determine the most query-like video segments in linear time complexity with respect to the number of parameters of the joint generative model. Our performance evaluation shows that our approach achieves an improvement in efficiency while maintaining high accuracy. In this work we make the following contributions:
(**i**) we investigate a parameter-based content representation for the purpose of endoscopic image and video linking,
(**ii**) we evaluate the qualities of accuracy and efficiency of our proposal in comparison to the state-of-the-art.

II. Related Work

This section only summarizes related works in the special field of image to video linking for endoscopic video archives. However, a comprehensive survey of content-based processing and analysis of endoscopic video is provided in the recent paper of Münzer et al. [3].

The problem of linking images with endoscopic video segments has been addressed already in the work of Roldan-Carlos et al. [4]. The authors utilize feature fusion based on *Color and Edge Directivity Descriptor* (CEDD) [5], *Color Correlograms* [6], and *Pyramid Histograms of Orientation Gradients* (PHOG) [7]. In addition to these global descriptors, the authors investigate the *SIMPLE* descriptor [8]. Beecks et al. [1] address the image-video-linking problem by utilizing low-dimensional position, color, and texture signatures in combination with a specific variant of the Signature Matching Distance [9]. Schoeffmann et al. [10] follow the signature-based approach and investigate other dissimilarity measures such as the Earth Mover's Distance [11] and the Signature Quadratic Form Distance [12].

As our proposal follows the aforementioned signature-based approaches [1], [10], we depict some endoscopic images together with their corresponding feature signatures – visualized by colored circles with diameters indicating their relevance – in Figure 1. The feature signatures, which are

Preprint of the paper to appear in the IEEE Proceedings of the Third International Conference on Multimedia Big Data(BigMM 2017), Laguna Hills, CA, USA, April 2017

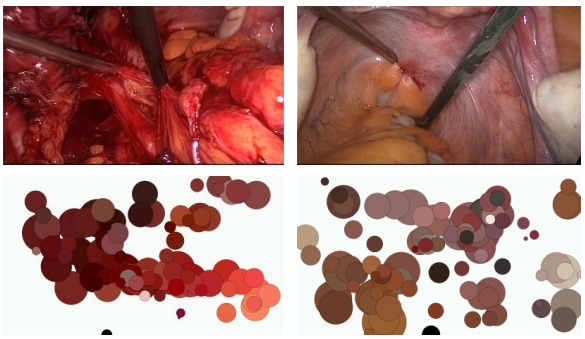


Figure 1. Visualizations of endoscopic images (top row) and their corresponding feature signatures (bottom row) [1]. The feature signatures are based on position, color, and texture information and adapt to individual visual content via object-specific feature quantizations.

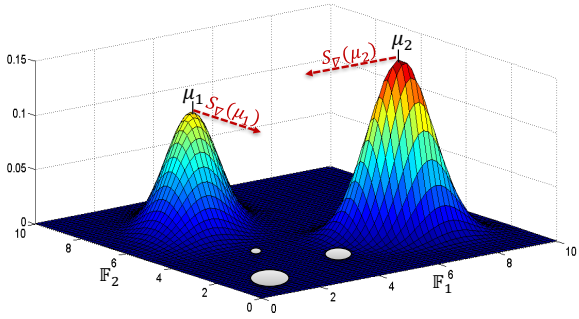


Figure 2. Illustration of the concept of a gradient-based signature. A Gaussian mixture model comprising two normal distributions that are centered at $\mu_1, \mu_2 \in \mathbb{F}$ is used to represent a feature signature $S \in \mathbb{R}^{\mathbb{F}}$, whose characteristic features are shown by gray circles. The red arrows $S_\nabla(\mu_1)$ and $S_\nabla(\mu_2)$ indicate the change of the parameters μ_1, μ_2 and are used to approximate the feature signature S.

based on position, color, and texture information, reflect the visual properties of the endoscopic images and are able to adapt to individual image/video content via object-specific feature quantizations.

Although the aforementioned approaches are able to accurately link endoscopic images to their corresponding video segments based on content-based visual features, they suffer from a high computational effort. Determining the most query-like video segments takes seconds up to minutes. In order to exploit the adaptability and flexibility of the feature signature model and reduce the computational effort, we propose to compactly approximate the content-based properties by means of gradient-based signatures. In this way, we advance from a feature-based content representation towards a parameter-based content representation.

III. Gradient-based Signatures for Endoscopic Images and Videos

The idea of gradient-based signatures is to approximate the content-based properties of endoscopic images and videos by means of a joint generative model. More specifically, the parameters of such a generative model are utilized in order to advance from a *feature-based content representation* to a *parameter-based content representation*. Thus, instead of expressing the visual properties explicitly by means of features in a feature space, as elucidated in the following Section III-A, we aim at reflecting the visual information implicitly through the parameters of a joint generative model, as proposed in Section III-B.

A. Feature-based Content Representation

The content-based properties of endoscopic images and videos are frequently represented by features $\{f_i\}_{i=1}^n \subset \mathbb{F}$ in a feature space $\mathbb{F}$. We abstract from numerical and categorical feature spaces and assume a general multi-dimensional Euclidean feature space capturing the images' visual characteristics in the remainder of this paper. Since each feature has a specific contribution to its corresponding image or video, a weighting is applied in order to model the features relevance. Mathematically, this weighting is defined as a function from the feature space $\mathbb{F}$ into the real numbers $\mathbb{R}$, which leads to the concept of a *feature signature* $S \in \mathbb{R}^{\mathbb{F}}$ as follows:

$$S : \mathbb{F} \to \mathbb{R} \text{ subject to } |\{f \in \mathbb{F} | S(f) \neq 0\}| < \infty. \quad (1)$$

A feature signature S assigns each feature $f \in \mathbb{F}$ from the feature space $\mathbb{F}$ a weight $S(f) \in \mathbb{R}$, where weights unequal to zero are designated for characteristic features which contribute to the corresponding image or video. Frequently but not necessarily, the weights of the characteristic features are positive. An example is given in Figure 1, where the characteristic features are depicted by circles accordingly.

By representing each endoscopic image or video, i.e., the frames of a video segment, by a feature signature, the individual content-based properties are aggregated into a finite feature-based content representation. Moreover, only those features with a weight unequal to zero have to be stored and indexed. In order to approximate these features and thus to further improve the storage complexity and processing efficiency, we propose the utilization of gradient-based signatures as shown in the following section.

B. Parameter-based Content Representation

As mentioned above, the idea of gradient-based signatures is to approximate the content-based properties of endoscopic images and videos, as expressed by the characteristic features and weights of the corresponding feature signatures, by means of the parameters of a joint generative model. Intuitively, we aim at adapting the parameters of the underlying generative model in order to better fit an individual feature signature and use these specific adaptations as parameter-based content representation.

This concept is illustrated in Figure 2, where we depict a generative model, i.e., a Gaussian mixture model,

comprising two normal distributions that are centered at $\mu_1, \mu_2 \in \mathbb{F}$ together with a feature signature $S \in \mathbb{R}^{\mathbb{F}}$. The characteristic features of S are shown by the gray circles. In order to better reflect the information that is expressed by the feature signature S, the parameters μ_1, μ_2 of the generative mode have to adapted accordingly and shifted into the directions indicated by the red arrows $S_\nabla(\mu_1)$ and $S_\nabla(\mu_2)$, respectively. In fact, these directions are then used as a parameter-based content representation of the feature signature S with respect to the underlying generative model.

This parameter-based content representation is mathematically defined by means of a *gradient-based signature* $S_\nabla \in \mathbb{R}^\theta$ from the model parameter space θ into the real numbers $\mathbb{R}$ as follows:

$$S_\nabla(\lambda) = \begin{cases} \nabla_\lambda \log \mathcal{L}(\theta|S) & , \lambda \in \theta \\ 0 & , otherwise \end{cases} \quad (2)$$

A gradient-based signature S_∇ approximates a given feature signature S via the gradients ∇_λ of the log-likelihood function $\log \mathcal{L}(\theta|S)$ of the model parameters $\lambda \in \theta$. Instead of storing and indexing all characteristic features with a weight unequal to zero, i.e., all features $f \in \mathbb{F}$ for which hold that $S(f) \neq 0$, only the change of the model parameters $\lambda \in \theta$ of the joint generative model have to be stored.

In case the generative model is a Gaussian mixture model $g \in \mathbb{R}^{\mathbb{F}}$ comprising n components with diagonal covariance matrices and parameters $\theta = \{\pi_k, \mu_{k_{[i]}}, \sigma_{k_{[i,j]}} | 1 \leq k \leq n \wedge 1 \leq i, j \leq d\}$ over a multi-dimensional Euclidean feature space $\mathbb{F} = \mathbb{R}^d$, the gradients $\nabla_{\mu_{k_{[i]}}} \log \mathcal{L}(\theta|S)$ with respect to the means $\mu_{k_{[i]}} \in \theta$ can be computed for a feature signature $S \in \mathbb{R}^{\mathbb{F}}$ by means of the following closed-form expression [2]:

$$\nabla_{\mu_{k_{[i]}}} \log \mathcal{L}(\theta|S) = \sum_{f \in \mathbb{F}} S(f) \cdot \frac{\pi_k \cdot \mathcal{N}(f|\mu_k, \Sigma_k)}{g(f|\theta)} \cdot \left(\frac{f_{[i]} - \mu_{k_{[i]}}}{\sigma^2_{k_{[i,i]}}} \right) \quad (3)$$

According to Equation 3, the computation time complexity needed to compute a single gradient $\nabla_{\mu_{k_{[i]}}} \log \mathcal{L}(\theta|S)$ is linear in the number of characteristic features of the corresponding feature signature S since the partial sum of non-characteristic features is zero.

The comparison of two gradient-based signatures $S_\nabla, S'_\nabla \in \mathbb{R}^\theta$ can be carried out along the joint model parameters $\lambda \in \theta$ in linear computation time complexity via the Minkowski distance as follows:

$$\mathrm{L}_p(S_\nabla, S'_\nabla) = \left(\sum_{\lambda \in \theta} |S_\nabla(\lambda) - S'_\nabla(\lambda)|^p \right)^{\frac{1}{p}} \quad (4)$$

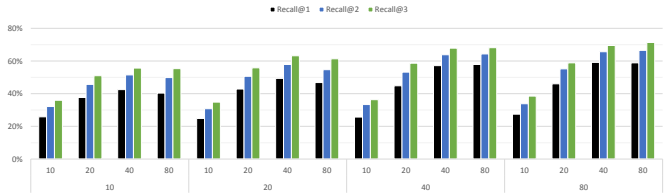
To sum up, gradient-based signatures approximate conventional feature signatures by means of a joint generative model. In comparison to the feature signature model, gradient-based signatures possess the following advantages: (i) By representing endoscopic images and videos by the model parameters, the storage and comparison complexities of gradient-based signatures is linear in the number of parameters and thus independent on the number of characteristic features of the feature signatures. (ii) Due to the parameter alignment for a given generative model, gradient-based signatures can be compared by conventional (dis)similarity measures for Euclidean spaces, such as the aforementioned Minkowski distances or cosine measures. We thus conclude that the utilization of gradient-based signatures allows us to efficiently link endoscopic images with their most similar video segments and evaluate the performance of our proposal in the following section.

IV. Performance Analysis

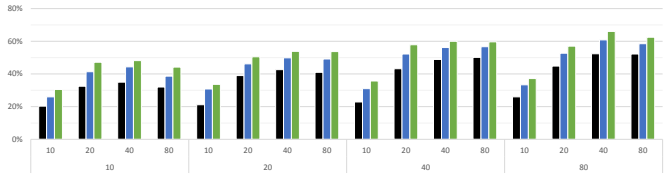
We evaluated the performance of gradient-based signatures in the context of linking endoscopic images to video segments. For this purpose, we used the same endoscopic video archive as provided in [4], [10], [1], which comprises more than 33 hours video content in full HD quality (1920x1080@25p) from 48 anonymized laparoscopic procedures. The data set additionally provides 600 selected endoscopic images, which are used as query images, with corresponding ground truth. The ground truth describes video segments that are related to each query image. Based on the ground truth, we measured the average recall@1, recall@2, and recall@3 values with respect to the query images and video segments in order to determine whether an approach is able to link an endoscopic image to its corresponding video segment.

We extracted feature signatures for the aforementioned endoscopic query images and video segments (5 fps) by first extracting local feature descriptors, denoted by PCT [9], which describe the relative spatial information, CIELAB color information, and coarseness and contrast information, and then clustering the extracted feature descriptors by the k-means algorithm. In this way, we obtained feature signatures with 10 up to 100 representatives over a 7-dimensional feature space $\mathbb{F} = \mathbb{R}^7$. In addition, we trained Gaussian mixture models with diagonal covariance matrices comprising 10 up to 100 components over the aforementioned feature space in order to compute gradient-based signatures with respect to the parameters $\theta_\mu = \{\mu_k\}_{k=1}^n$ of the mixture models.

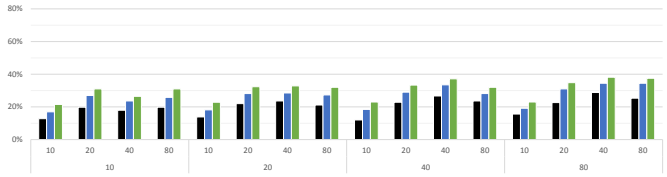
In the first series of experiments, we evaluated the performance of gradient-based signatures with respect to the quality of accuracy. To this end, we retrieved the most similar video segments for each endoscopic query image and averaged the corresponding recall@1, recall@2, and recall@3 values. The results for gradient-based signatures (GBS) in combination with different Minkowski distances L_p with $p \in \{1, 2, \infty\}$ are shown in Figure 3, while the results for the cosine similarity are shown in Figure 4. In these figures, the size of the query signatures and the size of the underlying Gaussian mixture models are both varied



(a) (GBS, L_1)

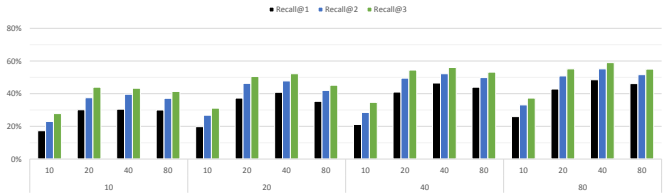


(b) (GBS, L_2)



(c) (GBS, L_∞)

Figure 3. Recall@1, recall@2, and recall@3 values in percentage for gradient-based signatures and different Minkowski distances as a function of the query signature size and the Gaussian mixture model size between 10, 20, 40 and 80, respectively.



(a) (GBS, cos)

Figure 4. Recall@1, recall@2, and recall@3 values in percentage for gradient-based signatures and the cosine similarity as a function of the query signature size and the Gaussian mixture model size between 10, 20, 40 and 80, respectively.

Table I
COMPARISON TO THE STATE-OF-THE-ART: RECALL VALUES AVERAGED OVER 600 QUERIES. THE QUERY SIGNATURE SIZE AND GAUSSIAN MIXTURE MODEL SIZE ARE GIVEN WHERE APPROPRIATE.

approach	**recall@1**	**recall@2**	**recall@3**
CEDD, PHOG [4]	78.3%	81.8%	84.2%
SIMPLE [4]	79.8%	83.3%	84.7%
SMD [1]	88.5% (50)	90.0% (80)	90.7% (80)
EMD [10]	66.5% (100)	71.3% (90)	74.3% (100)
SQFD [10]	82.2% (50)	85.8% (50)	87.5% (50)
GBS, L_1	62.5% (60/100)	69.0% (60/60)	72.0% (60/60)
GBS, L_2	54.0% (60/80)	61.3% (50/80)	66.0% (40/80)
GBS, L_∞	29.8% (50/80)	36.0% (40/70)	41.2% (40/70)
GBS, cos	50.2% (50/80)	56.0% (50/80)	59.3% (40/90)

between 10, 20, 40 and 80. As can be seen in the figures, an increase in the query signature size typically results in higher recall values across all evaluated recall levels. A similar behavior can be observed when increasing the number of components of the Gaussian mixture models, where more components in general lead to higher recall values and vice versa. With regard to the utilized (dis)similarity measures, the reported results clearly indicate that the Manhattan distance L_1 yields the highest recall values and thus the most accurate linking results. This is due to the perceptually uniformness of the inherent CIELAB color space that is used for encoding the color information. When utilizing gradient-based signatures in combination with the Manhattan distance (GBS, L_1) based on a query signature size of 60 and a Gaussian mixture model size of 100, the recall@1 value reaches 62.5%, while the recall@3 values exceeds 71%. In contrast, gradient-based signatures with the cosine similarity (GBS, cos) stay below a recall value of 60% across all evaluated recall levels. This leads to the observation that the individual differences along the model parameters carry the most useful information when determining the (dis)similarity between two gradient-based signatures. For this reason, the cosine similarity is outperformed by the Manhattan distance L_1 and the Euclidean distance L_2, since the latters' computations are attributed to the differences within the parameter changes.

To sum up, the reported recall values show that our proposal is able to link more than 70% of the endoscopic query images to their corresponding correct video segments. How this performance is related to the state-of-the-art is evaluated in the second series of experiments. Table I reports the average recall@1, recall@2, and recall@3 values for different state-of-the-art approaches. As can be seen from this table, signature-based approaches utilizing the Signature Matching Distance [1] are able to outperform feature-fusion-based approaches [4] using both global CEDD and PHOG descriptors as well as local SIMPLE descriptors. In general, the recall values of gradient-based signatures are below those of the other approaches since gradient-based signatures approximate the content-based information contained in the feature signatures, cf. Section III. In particular, in comparison to the signature-based approaches, the recall values of gradient-based signatures (GBS, L_1) are on average approximately 2% up to at most 21% lower than those of the Earth Mover's Distance (EMD) and Signature Matching Distance (SMD), respectively.

Nonetheless, it is worth noting that the major strengths of gradient-based signatures lie in an efficient means of comparison and thus in the trade-off between efficiency and accuracy. To this end, we empirically evaluated the performance in terms of average computation time that is required for linking an endoscopic image with its most similar video

Table II
AVERAGE COMPUTATION TIMES OF SIGNATURE-BASED APPROACHES IN SECONDS FOR DIFFERENT QUERY SIGNATURE SIZES.

query size	GBS	EMD [10]	SQFD [10]	SMD [1]
10	1.3	104.9	103.6	6.1
20	1.3	237.6	137.6	11.7
40	1.3	644.7	235.6	22.3
80	1.3	2033.8	546.4	42.9
average	1.3	755.2	255.8	20.8

segment. The computation times were measured without parallelization and without any additional indexing structure on a single-core CPU with 2.3 GHz and are averaged among different Gaussian mixture model sizes, where appropriate. The results are shown in Table II. Since the complexity of gradient-based signatures (GBS) solely depends on the number of model parameters, the computation time values are invariant with respect to the query signature size. In fact, gradient-based signatures need on average 1.3 seconds to link a specific endoscopic query image with its most similar video segment. In contrast to the other signature-based approaches whose average computation times range from 20 seconds (SMD) to more than 12 minutes (EMD), gradient-based signatures are on average more than 260 times faster than the investigated signature-based approaches.

To sum up, our performance analysis indicates that the problem of linking endoscopic images with video segments can be solved efficiently by approximating feature-based content representations via parameter-based content representations, i.e., by approximating feature signatures with gradient-based signatures. In this way, the performance in terms of accuracy slightly falls below that of state-of-the-art approaches, whereas the efficiency significantly improves. We thus conclude that gradient-based signatures provide a scalable solution for endoscopic image and video linking.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we have addressed the problem of linking endoscopic images with similar video segments in a content-based way without the utilization of additional semantic characteristics. To this end, we have investigated gradient-based signatures for endoscopic images and videos in order to advance from a feature-based content representation to a parameter-based content representation. Our performance analysis shows that our approach is on average more than 260 times faster than the state-of-the-art signature-based solutions while maintaining moderate-to-high accuracy.

In future work, we intend to increase the scalability of our approach by incorporating hashing techniques. Furthermore, we aim at applying data mining techniques for the investigation of large-scale endoscopic image and video archives.

ACKNOWLEDGMENTS

This work was supported by Universität Klagenfurt and Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF-20214 U. 3520/26336/38165.

REFERENCES

[1] C. Beecks, K. Schoeffmann, M. Lux, M. S. Uysal, and T. Seidl, "Endoscopic video retrieval: A signature-based approach for linking endoscopic images with video segments," in *Proceedings of the International Symposium on Multimedia*, 2015, pp. 33–38.

[2] C. Beecks, M. S. Uysal, J. Hermanns, and T. Seidl, "Gradient-based signatures for efficient similarity search in large-scale multimedia databases," in *Proc. of the Int. Conf. on Information and Knowledge Management*, 2015, pp. 1241–1250.

[3] B. Münzer, K. Schoeffmann, and L. Böszörmenyi, "Content-based proc. and analysis of endoscopic images and videos: A survey," *Multimedia Tools and Applications*, pp. 1–40, 2017.

[4] J. Roldan-Carlos, M. Lux, X. Giró i Nieto, P. Muñoz, and N. Anagnostopoulos, "Visual information retrieval in endoscopic video archives," in *Proceedings of the Int. Workshop on Content-Based Multimedia Indexing*, 2015, pp. 1–6.

[5] S. A. Chatzichristofis and Y. S. Boutalis, "CEDD: color and edge directivity descriptor: A compact descriptor for image indexing and retrieval," in *Proceedings of the International Conference on Computer Vision Systems*, 2008, pp. 312–322.

[6] J. Huang, R. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Image indexing using color correlograms," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 1997, pp. 762–768.

[7] A. Bosch, A. Zisserman, and X. Muñoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of the Int. Conf. on Image and Video Retrieval*, 2007, pp. 401–408.

[8] C. Iakovidou, N. Anagnostopoulos, A. C. Kapoutsis, Y. S. Boutalis, and S. A. Chatzichristofis, "Searching images with MPEG-7 (& mpeg-7-like) powered localized descriptors: The SIMPLE answer to effective content based image retrieval," in *Proceedings of the International Workshop on Content-Based Multimedia Indexing*, 2014, pp. 1–6.

[9] C. Beecks, S. Kirchhoff, and T. Seidl, "Signature matching distance for content-based image retrieval," in *Proceedings of the Int. Conference on Multimedia Retrieval*, 2013, pp. 41–48.

[10] K. Schoeffmann, C. Beecks, M. Lux, M. S. Uysal, and T. Seidl, "Content-based retrieval in videos from laparoscopic surgery," in *Proceedings of the SPIE Medical Imaging Conference*, 2016, pp. 97 861V–97 861V–10.

[11] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[12] C. Beecks, M. S. Uysal, and T. Seidl, "Signature quadratic form distance," in *Proceedings of the International Conference on Image and Video Retrieval*, 2010, pp. 438–445.