Proc. IEEE Int. Conf. on Image Processing (ICIP 2016), Phoenix, AZ, USA, Sept. 2016. DOI: 10.1109/ICIP.2016.7532791

VIDEO AESTHETIC QUALITY ASSESSMENT USING KERNEL SUPPORT VECTOR MACHINE WITH ISOTROPIC GAUSSIAN SAMPLE UNCERTAINTY (KSVM-IGSU)

Christos Tzelepis[⋆†] *Eftichia Mavridaki*[⋆] *Vasileios Mezaris*[⋆] *Ioannis Patras*[†]

[⋆] Information Technologies Institute (ITI), CERTH, Thermi 57001, Greece
[†]Queen Mary University of London, Mile end Campus, UK, E14NS
{tzelepis, eftichia88, bmezaris}@iti.gr, i.patras@qmul.ac.uk

ABSTRACT

In this paper we propose a video aesthetic quality assessment method that combines the representation of each video according to a set of photographic and cinematographic rules, with the use of a learning method that takes the video representation's uncertainty into consideration. Specifically, our method exploits the information derived from both low- and high-level analysis of video layout, leading to a photo- and motion-based video representation scheme. Subsequently, a kernel Support Vector Machine (SVM) extension, the KSVM-iGSU, is trained to classify the videos and retrieve those of high aesthetic value. Experimental results on our large dataset verify the effectiveness of the proposed method. We also make publicly available our dataset, in order to facilitate research in the area of video aesthetic quality assessment.

Index Terms— Video aesthetic quality assessment, rules of photography and cinematography, support vector machine, video representation uncertainty

1. INTRODUCTION

The assessment of the aesthetic quality of videos is a challenging field of research in today's digital world. In recent years, the amount of digital media has been growing significantly, and thus the development of effective aesthetic assessment methods is in great need in order to enhance multimedia content management in various applications, such as personal image collection management [1], food photo aesthetics assessment [2], and online fashion shopping photo assessment [3]. In the video domain, the automatic assessment of each video's aesthetic value could further improve the users' experience in multimedia content distribution channels, since videos could be retrieved or recommended by also taking their aesthetic quality into account.

In this paper we present a new method for Video Aesthetic Quality (VAQ) assessment, focusing primarily on the learning technique used for building a VAQ assessment system. First, we define a comprehensive representation scheme by exploiting photo- and motion-based features, motivated by photography and cinematography rules. Then, we introduce the use of a sophisticated Support Vector Machine (SVM) extension, such that the uncertainty that is encapsulated in the representation of the input videos is taken into consideration during training. For the purpose of evaluation, we treat the VAQ assessment task as a retrieval problem, since a VAQ system naturally needs not only to make a binary decision on whether a given video is of high-aesthetic quality or not, but also to rank the aesthetic quality of videos within a dataset, such that videos of higher aesthetic value can be ranked higher. Furthermore, we make publicly available a large video dataset with ground-truth aesthetic quality annotations, in order to facilitate further research in this field.

The remainder of the paper is organized as follows. In Section 2 we review the related work. In Section 3 we describe the video features. In Section 4 we introduce the use of KSVM-iGSU in the VAQ assessment problem. In Section 5 we present our new public dataset. In Section 6, we experimentally validate the proposed method and discuss the results, and finally, conclusions are drawn in Section 7.

2. RELATED WORK

Only a few methods for video aesthetic quality assessment have been proposed so far. The first methods in this domain tried to estimate the video aesthetic value by extracting mostly low-level features from video frames. For instance, in [4], a set of low-level features, such as sharpness, colorfulness, luminance and blockiness quality, and a few motion features, are extracted. Then, an SVM using the Radial Basis Function (RBF) kernel is trained for assessing the aesthetic quality of videos. In [5], the authors treat the video as a sequence of still images to which they apply a set of visual-based features together with two additional motion-based features, i.e., the length of subject region motion and motion stability, so as to distinguish professional videos from amateurish ones. They also applied a set of learning approaches, such as kernel SVM, Bayesian classification, and Gentle AdaBoost.

This work was supported by the EU's Horizon 2020 programme under grant agreements H2020-687786 InVID and H2020-693092 MOVING.

Proc. IEEE Int. Conf. on Image Processing (ICIP 2016), Phoenix, AZ, USA, Sept. 2016. DOI: 10.1109/ICIP.2016.7532791

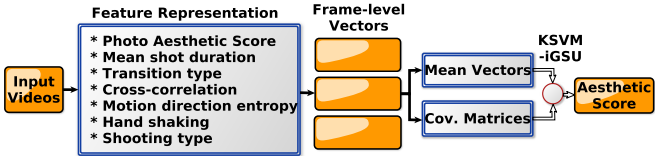

Fig. 1: Proposed VAQ assessment approach.

A more elaborate method that introduces a set of features ranging from low- and mid-level attributes to high-level style descriptors, combined with a kernel SVM learning stage, is presented in [6]. In [7], an RBF SVM is applied to a set of "semantically independent" features, such as camera motion and stabilization, and frame composition, along with a set of "semantically dependent" features, such as motion direction entropy, color saturation, and lightness. Semantic dependency of a feature refers to whether this feature relates or not to the semantic content of each frame. Moreover, in [8], low- and high-level visual and motion features are extracted at cell-, frame-, and shot-level and a Low Rank Late Fusion (LRLF) scheme is used for fusing the scores produced by a set of SVMs, each of which was trained with one specific aesthetic feature. More motion features are introduced in [9], where the authors evaluate the effectiveness of motion space, motion direction entropy and hand shaking (i.e., camera stabilization) on VAQ assessment tasks. They also use naive Bayesian, SVM, and AdaBoost classification techniques. Finally, in [10], a variety of aesthetic-related features for video are designed, such as visual continuity and shot length, and their performance in retrieving professional videos in conjunction with a kernel SVM classifier is examined.

The video aesthetic assessment techniques in the literature are typically evaluated on different datasets: these include the NHK "Where is beauty?" dataset [11], the Telefonica dataset [9], the dataset of [10], etc. The NHK dataset consists of 1000 professionally-produced video segments that last about 1 min each. The Telefonica dataset consists of 160 short consumer videos (each being 11 to 60 seconds long), for which the mean opinion scores (MOS) in terms of aesthetic quality are provided. The dataset of [10] includes 1000 professionally-generated videos, and 1000 amateur videos. Most of these datasets are not publicly available.

3. VIDEO'S AESTHETIC FEATURES

Since video is typically treated as a sequence of still images that gives the impression of motion, both the visual and motion modalities need to be exploited in order to effectively evaluate its aesthetic quality. For the purpose of the present study, each video is described according to a set of rules borrowed from photography and cinematography. Initially, each video is divided into its shots using the shot detection method of [12]. Then, for each video, we estimate the mean duration of its shots, and, considering that the shot transitions can be either abrupt or gradual, we estimate for each of these transition types their duration as a percentage of the whole video's duration. This results in a 3-element video-level vector.

Subsequently, one keyframe per second is extracted from the original raw video sequence (irrespective of shot boundaries), and photo- and motion- based features are extracted for each one of them. Photo-based features include the simplicity, colorfulness, sharpness, pattern and overall aesthetic quality values, which are extracted based on the still-image aesthetic quality assessment method proposed in [13]. Motion-based features, adopted from [9], include: a) a measure of similarity between successive frames (cross-correlation between these frames), b) a measure of the diversity of motion directions (motion direction entropy), c) a measure of the stability of the camera during the capturing process (hand-shaking), and d) a measure which can distinguish the difference between three categories of shots: focused shots, panorama shots and static shots (shooting type). The above result in a 44-element keyframe-level feature vector. Concatenating with it the video-level feature vector, we end up with a 47-element vector as the final representation for each keyframe.

4. KERNEL SVM WITH ISOTROPIC GAUSSIAN SAMPLING UNCERTAINTY (KSVM-IGSU)

A challenge in the video aesthetic quality assessment problem, similarly to many video classification tasks, is that video representation techniques usually introduce uncertainty in the input that is subsequently fed to the classifiers. Thus, uncertainty needs to be taken into consideration during classifier training. The kernel SVM with Isotropic Gaussian Sample Uncertainty (KSVM-iGSU), proposed in [14], is an extension of the standard kernel SVM that exploits the uncertainty of input data in order to achieve better classification results. The uncertainty of the i-th input example is modeled as an isotropic Gaussian distribution with given mean vector $\mathbf{x}_i \in \mathbb{R}^n$ and an isotropic covariance matrix, i.e. a scalar multiple of the identity matrix, $\Sigma_i = \sigma_i^2 I_n \in \mathbb{S}_{++}^n$[1], where n denotes the dimensionality of the input feature space.

The optimization problem of KSVM-iGSU can be cast as a variational calculus problem of finding the function f that minimizes the functional $\Phi[f]$, i.e., $\min_{f\in\mathcal{H}} \Phi[f]$, where the functional $\Phi[f]$ is given by

$$\Phi[f] = \frac{1}{2}\lambda\|f\|_{\mathcal{H}}^2 + \frac{1}{\ell}\sum_{i=1}^{\ell}\left[\frac{y_i - f(\mathbf{x}_i) - b}{2}\left(\operatorname{erf}\left(\frac{y_i - f(\mathbf{x}_i) - b}{\sqrt{2\sigma_i^2\|f\|_{\mathcal{H}}^2}}\right) + y_i\right) + \frac{\sqrt{\sigma_i^2\|f\|_{\mathcal{H}}^2}}{\sqrt{2\pi}}\exp\left(\frac{(y_i - f(\mathbf{x}_i) - b)^2}{2\sigma_i^2\|f\|_{\mathcal{H}}^2}\right)\right], \quad (1)$$

[1] $\mathbb{S}_{++}^n$ denotes the convex cone of all symmetric positive definite $n \times n$ matrices with entries in $\mathbb{R}$. I_n denotes the identity matrix of order n.

Proc. IEEE Int. Conf. on Image Processing (ICIP 2016), Phoenix, AZ, USA, Sept. 2016. DOI: 10.1109/ICIP.2016.7532791

where $\lambda = \frac{1}{C}$ is a regularization parameter, C is the tradeoff parameter of standard SVM, and f belongs to a Reproducing Kernel Hilbert Space (RKHS), $\mathcal{H}$, with associated kernel k. Using a generalized semi-parametric version [15] of the representer theorem [16], it can be shown that the minimizer of the above functional admits a solution of the form

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i k(\mathbf{x}, \mathbf{x}_i) - b, \tag{2}$$

where $b \in \mathbb{R}$, $\alpha_i \in \mathbb{R}$, $\forall i$.

We define the kernel matrix K as the symmetric positive definite $\ell \times \ell$ matrix given as $K = (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^{\ell}$. If we set $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_\ell)^\top$ and let $\mathbf{K}_i$ denote the i-th column of the kernel matrix K, the objective function of KSVM-iGSU, $\mathcal{J}_{\mathcal{H}}: \mathbb{R}^\ell \times \mathbb{R} \to \mathbb{R}$, can be rewritten as follows

$$\mathcal{J}_{\mathcal{H}}(\boldsymbol{\alpha}, b) = \frac{1}{2}\lambda \boldsymbol{\alpha}^\top K \boldsymbol{\alpha} + \frac{1}{\ell}\sum_{i=1}^{\ell}\left[\frac{y_i - \mathbf{K}_i \cdot \boldsymbol{\alpha} - b}{2}\left(\operatorname{erf}\left(\frac{y_i - \mathbf{K}_i \cdot \boldsymbol{\alpha} - b}{\sqrt{2\sigma_i^2 \boldsymbol{\alpha}^\top K \boldsymbol{\alpha}}}\right) + y_i\right) + \frac{\sqrt{\sigma_i^2 \boldsymbol{\alpha}^\top K \boldsymbol{\alpha}}}{\sqrt{2\pi}} \exp\left(-\frac{(y_i - \mathbf{K}_i \cdot \boldsymbol{\alpha} - b)^2}{2\sigma_i^2 \boldsymbol{\alpha}^\top K \boldsymbol{\alpha}}\right)\right], \tag{3}$$

where the sum above expresses the total loss. We (jointly) minimize the above convex objective function with respect to $\boldsymbol{\alpha}$, b using the Limited-memory BFGS (L-BFGS) algorithm [17]. L-BFGS is a quasi-Newton optimization algorithm that approximates the Broyden-Fletcher-Goldfarb-Shanno (BFGS) [18] algorithm using a limited amount of computer memory. Since $J_{\mathcal{H}}$ is a convex function on $\mathbb{R}^\ell \times \mathbb{R}$ (see [14]), L-BFGS leads to a global optimal solution; that is, at a pair $(\boldsymbol{\alpha}, b)$ such that the decision function given in the form of (2) minimizes the functional (1).

5. PUBLIC DATASET FOR VAQ ASSESSMENT

Existing datasets containing only very short video segments (e.g., < 60 seconds), extracted from professionally-produced videos (e.g., Hollywood movies), are not sufficiently representative of the user-generated content found in social platforms such as YouTube. Datasets used in previous works typically consist of short videos where only a very few number of shots can be detected. For instance, most videos in the Telefonica dataset [9] are made of a single shot. As a result, video-structure features cannot be assessed reliably. Moreover, video datasets derived from professionally-generated content are not suitable for training and evaluating methods for the assessment of user-generated videos, which is where VAQ assessment is most useful and needed in practice. The NHK dataset [11] contains such videos that either are derived from Hollywood movies or have been shot using professional equipment and stabilized cameras.

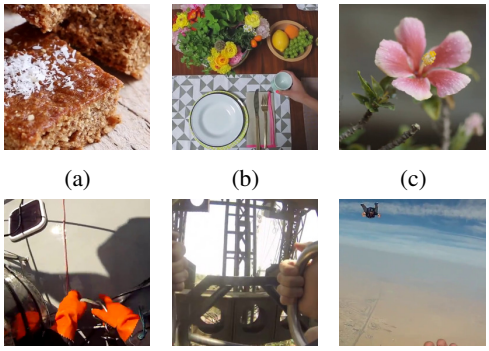

Fig. 2: Indicative examples of videos of high (a,b,c) and low (d,e,f) aesthetic value, available in our dataset.

We aim to perform video aesthetic quality assessment under conditions that are as close to real-life scenarios as possible. For this reason, we introduce a new video dataset that consists of user-created videos, capturing moments of everyday life, such as excursions, school concerts, and training processes. We downloaded from YouTube 700 videos covering a variety of categories, such as outdoor activities, do it yourself videos, make up tutorials, lectures, and home-made videos, licensed under Creative Commons Attribution [19]. The duration of each of these videos ranges from 1 to 6 minutes.

Subsequently, we conducted an annotation process that involved 12 annotators watching and evaluating the aesthetic value of each of these videos by assigning binary aesthetic quality ratings; 1 being assigned to videos of high aesthetic quality and 0 to videos of low aesthetic quality. Each video was assessed by 5 annotators. Before the annotation process, the annotators watched some indicative examples of videos of high and low aesthetic quality, and were instructed to remain as uninfluenced as possible by the video's semantics. The final aesthetic score of each annotated video was calculated as the median score of the annotators' individual scores.

As a result of the annotation process, 350 videos are rated as being of high aesthetic quality and another 350 as being of low aesthetic quality. Indicative frames of such videos are shown in Fig. 2. We call this dataset CERTH-ITI-VAQ700 and we make it publicly available for research purposes[2].

6. EXPERIMENTAL RESULTS AND DISCUSSION

6.1. Experimental Setup

In our experiments, the dataset of Section 5 is randomly split into a training subset (50%) and an evaluation subset (50%), each maintaining a positive-negative ratio of 1 : 1. That is,

[2]Our video aesthetic quality assessment dataset and the corresponding ground-truth annotation are publicly available at http://mklab.iti.gr/project/certh-iti-vaq700-dataset.

Proc. IEEE Int. Conf. on Image Processing (ICIP 2016), Phoenix, AZ, USA, Sept. 2016. DOI: 10.1109/ICIP.2016.7532791

each of the training and evaluation subsets includes 175 positive (high aesthetic) and 175 negative (low aesthetic) video examples. As discussed in Section 3, for video representation, 1 keyframe per second was extracted at regular time intervals from each video, and each keyframe was represented using the proposed photo- and motion-based features.

The aforementioned keyframe-level video representations can be seen as observations of the input Gaussian distributions that describe the training videos. That is, let $\mathcal{X}$ be a set of ℓ annotated random vectors representing the video-level feature vectors. We assume that each random vector is distributed normally; i.e., for the random vector representing the i-th video, $\mathbf{X}_i$, we have $\mathbf{X}_i \sim \mathcal{N}(\mathbf{x}_i, \Sigma_i)$. Also, for each random vector $\mathbf{X}_i$, a number, N_i, of observations, $\{\mathbf{x}_i^t \in \mathbb{R}^n : t = 1, \ldots, N_i\}$ are available; these are the keyframe-level feature vectors that have been computed. Then, the mean vector and the covariance matrix of $\mathbf{X}_i$ are computed respectively as follows

$$\mathbf{x}_i = \frac{\sum_{t=1}^{N_i} \mathbf{x}_i^t}{N_i}, \quad \Sigma_i = \frac{\sum_{t=1}^{N_i} (\mathbf{x}_i^t - \mathbf{x}_i)(\mathbf{x}_i^t - \mathbf{x}_i)^\top}{N_i - 1} \tag{4}$$

Now, due to the assumption for isotropic covariance matrices, we approximate the above covariance matrices as multiples of the identity matrix, i.e. $\widehat{\Sigma_i} = \sigma_i^2 I_n$. As discussed in [14], it suffices to set σ_i^2 equal to the mean value of the elements of the main diagonal of Σ_i.

Since the problem of video aesthetic quality assessment can be naturally seen as a retrieval application, where a user queries for videos of high aesthetic quality within a dataset, for assessing the performance of our method we use retrieval-oriented evaluation measures: the average precision (AP) [20], as well as the precision at depth n (where $n \in \{5, 10, 15, 20\}$) and the accuracy, which are measures that are typically used in the VAQ assessment literature [5, 7, 9, 10, 13].

6.2. Experimental Results

The proposed KSVM-iGSU-based method is tested and compared to the standard kernel SVM (KSVM). KSVM is the state-of-the-art classifier for the problem of VAQ assessment [4, 7, 9, 10, 13]. For both KSVM-iGSU and KSVM, the radial basis function (RBF) kernel was used. Training parameters C, γ were obtained via a 3-fold cross-validation procedure (grid search) with C being searched in the range $\{2^{-4}, 2^{-3}, \ldots, 2^6, 2^7\}$ and γ in the range $\{2^{-7}, 2^{-6}, \ldots, 2^3, 2^4\}$. Each of the above experiments was repeated 10 times using different random training/evaluation subsets, similarly to [13].

Table 1 shows the average performance of KSVM-iGSU compared to the standard KSVM in terms of precision for the top-n retrieved videos, where $n = 5, 10, 15, 20$. We see that the proposed VAQ assessment method leads to considerably better results in terms of retrieval precision. Furthermore, the mean values of accuracy (AC) and average precision (AP) are also reported for 10 repetitions of the experiment. We see that, in terms of accuracy, the proposed method slightly outperforms the state-of-the-art KSVM (0.6814 over 0.6787, respectively), but in terms of average precision, which is a more meaningful measure for retrieval tasks, our method reaches 0.6997 as compared to KSVM's 0.6167, leading to a 13.45% relative boost. Finally, Fig. 3 shows the recall-precision curves of the proposed method (KSVM-iGSU) and standard KSVM.

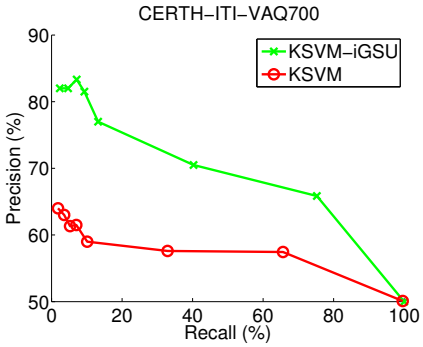

Fig. 3: Precision-Recall curves for the proposed VAQ assessment method (KSVM-iGSU) compared to the state-of-the-art KSVM approach using CERTH-ITI-VAQ700 dataset.

Table 1: Performance of the proposed method (using KSVM-iGSU) compared to the standard KSVM in terms of precision at top-n retrieved videos ($n = 5, 10, 15, 20$), accuracy (AC), and average precision (AP) using CERTH-ITI-VAQ700 dataset.

	KSVM (as in [7,9,10])	**KSVM-iGSU (Proposed)**
P@5	0.6400	**0.8200**
P@10	0.6300	**0.8200**
P@15	0.6133	**0.8333**
P@20	0.6150	**0.8150**
AC	0.6787	**0.6814**
AP	0.6167	**0.6997**

7. CONCLUSION

In this paper we proposed a new video aesthetic quality assessment method that combines a comprehensive set of video features with a new learning approach, which takes the video representation's uncertainty into consideration. We also make publicly available our VAQ assessment dataset in order to facilitate further research in this area. Experimental results of our approach demonstrate considerable performance improvement in comparison to the state-of-the-art learning methods used for video aesthetic quality assessment.

8. REFERENCES

[1] F. Simond, N. Arvanitopoulos Darginis, and S. Süsstrunk, "Image aesthetics depends on context," in *Proc. of Int. Conf. on Image Processing (ICIP)*, 2015.

[2] Y. Li and A. Sheopuri, "Applying image analysis to assess food aesthetics and uniqueness," in *Proc. of Int. Conf. on Image Processing (ICIP)*. IEEE, 2015, pp. 311–314.

[3] J. Wang and J. Allebach, "Automatic assessment of online fashion shopping photo aesthetic quality," in *Proc. of Int. Conf. on Image Processing (ICIP)*. IEEE, 2015, pp. 2915–2919.

[4] A. K. Moorthy, P. Obrador, and N. Oliver, "Towards computational models of the visual aesthetic appeal of consumer videos," in *Proc. of the 11th European Conference on Computer Vision (ECCV), Heraklion, Crete, Greece, 2010*, pp. 1–14. Springer, 2010.

[5] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proc. of the 10th European Conference on Computer Vision (ECCV), Marseille, France*, pp. 386–399. Springer, 2008.

[6] Y. Wang, Q. Dai, R. Feng, and Y.-G. Jiang, "Beauty is here: Evaluating aesthetics in videos using multimodal features and free training data," in *Proc. of the 21st ACM Int. Conf. on Multimedia*. ACM, 2013, pp. 369–372.

[7] C.-Y. Yang, H.-H. Yeh, and C.-S. Chen, "Video aesthetic quality assessment by combining semantically independent and dependent features," in *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 1165–1168.

[8] S. Bhattacharya, B. Nojavanasghari, T. Chen, D. Liu, S.-F. Chang, and M. Shah, "Towards a comprehensive computational model foraesthetic assessment of videos," in *Proc. of the 21st Int. Conf. on Multimedia*. ACM, 2013, pp. 361–364.

[9] H.-H. Yeh, C.-Y. Yang, M.-S. Lee, and C.-S. Chen, "Video aesthetic quality assessment by temporal integration of photo-and motion-based features," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1944–1957, 2013.

[10] Y. Niu and F. Liu, "What makes a professional video? a computational aesthetics approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 7, pp. 1037–1049, 2012.

[11] Grand Challenge at ACM Multimedia Conf. (MM'13), "NHK Where is beauty?," http://acmmm13.org/submissions/call-for-multimedia-grand-challenge-solutions/task-where-is-beauty/, Barcelona, Spain, October 2013.

[12] E. Apostolidis and V. Mezaris, "Fast shot segmentation combining global and local visual descriptors," in *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 6583–6587.

[13] E. Mavridaki and V. Mezaris, "A comprehensive aesthetic quality assessment method for natural images using basic rules of photography," in *Proc. of Int. Conf. on Image Processing (ICIP)*. IEEE, 2015, pp. 887–891.

[14] C. Tzelepis, V. Mezaris, and I. Patras, "Video event detection using kernel support vector machine with isotropic gaussian sample uncertainty (KSVM-iGSU)," in *Proc. of the 22nd Int. Conf. on MultiMedia Modeling (MMM), Miami, FL, USA*. Springer, 2016, pp. 3–15.

[15] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Proc. of 14th Annual Conf. on Computational Learning Theory*, 2001, pp. 416–426.

[16] G. Kimeldorf and G. Wahba, "Some results on Tchebycheffian spline functions," *Journal of Mathematical Analysis and Applications*, vol. 33, no. 1, pp. 82–95, 1971.

[17] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical programming*, vol. 45, no. 1-3, pp. 503–528, 1989.

[18] C. G. Broyden, "The convergence of a class of double rank minimization algorithms 1. general considerations," *Journal of Applied Mathematics*, vol. 6, no. 1, pp. 76–90, 1970.

[19] "Creative commons attribution license," https://creativecommons.org/, December 2015.

[20] S. Robertson, "A new interpretation of average precision," in *Proc. of the 31st Annual Int. ACM SIGIR conf. on Research and Development in Information Retrieval*. ACM, 2008, pp. 689–690.