# Learning Memorability Preserving Subspace for Predicting Media Memorability

Yang Liu[1,2], Zhonglei Gu[1], Tobey H. Ko[3]
[1]Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, P.R. China
[2]HKBU Institute of Research and Continuing Education, Shenzhen, P.R. China
[3]Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Hong Kong SAR, P.R. China
csygliu@comp.hkbu.edu.hk,cszlgu@comp.hkbu.edu.hk,tobeyko@hku.hk

## ABSTRACT

This paper describes our approach designed for the MediaEval 2018 Predicting Media Memorability Task. First, a subspace learning method called Memorability Preserving Embedding (MPE) is proposed to learn discriminative subspace from the original feature space according to the memorability scores. Then the Support Vector Regressor (SVR) is applied to the learned subspace for memorability prediction. The prediction performance demonstrates that SVR can achieve good performance even in a very low-dimensional subspace, which implies that the subspace learned by the MPE is capable of preserving important memorability information. Moreover, the results indicate that the short-term memorability is more predictable than the long-term memorability.

## 1 INTRODUCTION

Predicting media memorability plays a key role in many real-world applications such as media retrieval and recommendation, and has attracted much attention recently [1, 4, 6, 9–12, 14]. The *MediaEval 2018 Predicting Media Memorability Task* aims to seek solutions to the problem of predicting how memorable a video will be [3]. Specifically, given a set of training video data (each data sample is associated with its visual features and the corresponding memorability s-core), the participants are asked to build a model using the training data and utilize the trained model to predict the memorability score of test data.

Images and videos often have very high dimensionality, which brings computational challenges to the analysis tasks. To solve the memorability prediction task in an efficient way, in this paper, we propose a supervised subspace learning method called Memorability Preserving Embedding (MPE). The motivation of designing such a subspace learning method for the task rather than directly performing the prediction is that we believe most of the discriminative information of the high-dimensional media data is actually embedded in a relatively low-dimensional subspace and discovering such a subspace could enhance the performance of prediction. Therefore, the proposed MPE aims to learn a transformation matrix to project the high-dimensional training data to a low-dimensional subspace, in which the memorability information and manifold structure of the dataset are well preserved. In the test stage, we use the learned transformation matrix to map the test data to the subspace, and apply a Support Vector Regressor (SVR) [13] to the subspace for final memorability prediction.

## 2 MEMORABILITY PRESERVING EMBEDDING

Given the training set $\mathcal{X} = \{(\mathbf{x}_1, l_1), (\mathbf{x}_2, l_2), ..., (\mathbf{x}_n, l_n)\}$, with $\mathbf{x}_i \in \mathbb{R}^D$ $(i = 1, \cdots, n)$ being the visual feature vector of the $i$-th video and $l_i \in [0, 1]$ being the corresponding memorability score, MPE aims to learn a $D \times d$ transformation matrix $\mathbf{W}$ to map $\mathbf{x}_i$ $(i = 1, \cdots, n)$ to a low-dimensional subspace, where the memorability information and manifold structure of the dataset can be well preserved. To achieve this goal, MPE optimizes the following objective function:

$$\mathbf{W} = \arg\min_{\mathbf{W}} \sum_{i,j=1}^{n} \|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|^2 \cdot \big(\alpha S_{ij} + (1-\alpha) N_{ij}\big), \quad (1)$$

where $S_{ij} = exp(-(l_i - l_j)^2/2\sigma^2)$ measures the similarity between the memorability score of $\mathbf{x}_i$ and that of $\mathbf{x}_j$, $N_{ij} = exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma^2)$ measures the closeness between $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\alpha \in [0, 1]$ is the parameter balancing the memorability information and the manifold structure.

Eq. (1) could be equivalently rewritten as follows:

$$\mathbf{W} = \arg\min_{\mathbf{W}} tr(\mathbf{W}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{W}), \quad (2)$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \mathbb{R}^{D \times n}$ is the data matrix, $\mathbf{L} = \mathbf{D} - \mathbf{A}$ is the $n \times n$ Laplacian matrix [7], and $\mathbf{D}$ is a diagonal matrix defined as $D_{ii} = \sum_{j=1}^{n} A_{ij}$ $(i = 1, ..., n)$, where $A_{ij} = \alpha S_{ij} + (1 - \alpha) N_{ij}$. Then the optimal $\mathbf{W}$ can be obtained by finding the eigenvectors corresponding to the smallest eigenvalues of the following eigen-decomposition problem:

$$\mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{w} = \lambda \mathbf{w}. \quad (3)$$

After obtaining $\mathbf{W}$, for each high-dimensional data sample $\mathbf{x}_i$ in the development and test sets, we can obtain its low-dimensional representation by $\mathbf{y}_i = \mathbf{W}^T \mathbf{x}_i$. Then we apply SVR to $\mathbf{y}_i$ for memorability prediction.

*MediaEval'18, 29-31 October 2018, Sophia Antipolis, France*

**Table 1: The performance (in terms of Spearman Correlation and MSE) of our approach on the test set of MediaEval 2018 Predicting Media Memorability Task.**

| | | Run1<br>($d = 4$) | Run2<br>($d = 5$) | Run3<br>($d = 9$) | Run4<br>($d = 10$) |
|---|---|---|---|---|---|
| Spearman | Long | 0.0774 | 0.0962 | 0.0647 | 0.0634 |
| | Short | 0.1332 | 0.1268 | 0.0656 | 0.0717 |
| MSE | Long | 0.0214 | 0.0214 | 0.0213 | 0.0213 |
| | Short | 0.0082 | 0.0080 | 0.0078 | 0.0079 |

## 3 RESULTS AND ANALYSIS

In this section, we report our experimental results on the MediaEval 2018 Predicting Media Memorability Task [3]. Specifically, we participate in two subtasks: 1) short-term memorability subtask and 2) long-term memorability subtask.

We use both video specialized features and image features, which are provided by the task, to construct the original feature space. For the video features, we use the 101-D C3D feature vector. For the image features, we use the 122-D local binary pattern (LBP) feature vector and the 768-D color histogram feature vector. We select these features as they have demonstrated good performance in visual analysis tasks [5, 8, 15]. For each video, the first, the median, and the last frames are selected as the representatives of the video, so the total dimension of the original feature space is $D = 101 + 3 \times (122 + 768) = 2771$.

We use all 8000 video data samples in the development set for training. Before subspace learning, we normalize the values of different features to $[0, 1]$. For the MPE method, we set $\alpha = 0.5$ and $\sigma = 1$.

- For Run 1, we set the reduced dimension $d = 4$. Then we learn the $D \times d$ (i.e., $2771 \times 4$ in this case) transformation matrix $\mathbf{W}$ via MPE using the development set, and utilize $\mathbf{W}$ to map both development and test data onto the 4-D subspace. Finally, we train the $\nu$-SVR [13] using the development set in the 4-D subspace and employ the trained $\nu$-SVR model to predict the memorability score of the test data in the same subspace. We use the RBF kernel and set $\nu = 0.5$ and $\gamma = 1/D$ [2].
- For Run 2, we set the reduced dimension $d = 5$.
- For Run 3, we set the reduced dimension $d = 9$.
- For Run 4, we set the reduced dimension $d = 10$. The remaining procedure and the parameter setting in Runs 2, 3, and 4 are the same as those in Run 1.

Table 1 shows the performance (in terms of Spearman Correlation and MSE) of our approach. From the results, we have several observations. First, we observe that the results (both Spearman and MSE) on the short-term subtask are better than those on the long-term subtask, which indicates that the short-term memorability is more predictable than the long-term memorability. Besides, by comparing the MSE of runs 1 and 2 ($d = 4, 5$) and that of runs 3 and 4 ($d = 9, 10$), we notice that runs 1 and 2 are better than runs 3 and 4 in terms of Spearman, and are comparable in terms of MSE. This fact may imply that most of the discriminative information is embedded in a very low-dimensional subspace and increasing more dimensions may not necessarily improve the performance.

**Table 2: The performance (in terms of Spearman Correlation and MSE) of our approach on the development set of MediaEval 2018 Predicting Media Memorability Task.**

| | | $d = 4$ | $d = 5$ | $d = 9$ | $d = 10$ | $D$ |
|---|---|---|---|---|---|---|
| Spearman | Long | 0.1422 | 0.1514 | 0.1654 | 0.1675 | 0.1414 |
| | Short | 0.3047 | 0.3059 | 0.3065 | 0.3070 | 0.2946 |
| MSE | Long | 0.0212 | 0.0212 | 0.0211 | 0.0210 | 0.0211 |
| | Short | 0.0061 | 0.0061 | 0.0061 | 0.0061 | 0.0062 |

To further validate the effectiveness of subspace learning, we compare the performance of SVR on the learned subspace and that on the original 2771-D space using the development set. We use 5-fold cross validation and average the results. The Spearman coefficient and MSE in Table 2 show that the performance on the original space is slightly worse than that on learned subspaces, supporting our assumption that the original high-dimensional space may contain redundant or even noisy information, and reducing the dimensionality with supervised information could improve the subsequent learning performance. However, the results in terms of Spearman coefficient is far from satisfactory. The reason might be that MPE is a linear mapping method, which is not sufficient to capture the complex discriminant information embedded in the high-dimensional feature space. This motivates us to consider extending our method to the nonlinear case to improve the performance.

## 4 CONCLUSION

This paper describes our approach designed for memorability prediction. A subspace learning method, MPE, is proposed to learn the subspace that preserves the memorability information. After that, SVR is utilized for memorability prediction in the learned subspace. The results on the MediaEval 2018 Predicting Media Memorability Task validate the effectiveness of our approach. Our future work will focus on exploring the physical meaning of the learned subspace, as this could improve the interpretability of our approach. Moreover, we plan to generalize our method to nonlinear scenario to enhance its data representation ability.

## ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61503317 and in part by the General Research Fund (GRF) from the Research Grant Council (RGC) of Hong Kong SAR under Project HKBU12202417.

## REFERENCES

[1] Y. Baveye, R. Cohendet, M. Perreira Da Silva, and P. Le Callet. 2016. Deep Learning for Image Memorability Prediction: The Emotional Bias. In *Proceedings of the 24th ACM International Conference on Multimedia (MM '16)*. ACM, New York, NY, USA, 491–495.

[2] C.-C. Chang and C.-J. Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2 (2011), 27:1–27:27. Issue 3.

[3] R. Cohendet, C.-H. Demarty, N. Q. K. Duong, M. Sjoberg, B. Ionescu, and T.-T. Do. MediaEval 2018: Predicting Media Memorability. In *Proceedings of the MediaEval 2018 Workshop*. CEUR-WS, Sophia Antipolis, France, 29–31 October, 2018.

[4] R. Cohendet, K. Yadati, N. Q. K. Duong, and C.-H. Demarty. 2018. Annotating, Understanding, and Predicting Long-term Video Memorability. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval (ICMR '18)*. ACM, New York, NY, USA, 178–186.

[5] A. M. Ferman, A. M. Tekalp, and R. Mehrotra. 2002. Robust color histogram descriptors for video segment retrieval and identification. *IEEE Transactions on Image Processing* 11, 5 (2002), 497–508.

[6] J. Han, C. Chen, L. Shao, X. Hu, J. Han, and T. Liu. 2015. Learning Computational Models of Video Memorability from fMRI Brain Imaging. *IEEE Transactions on Cybernetics* 45, 8 (Aug 2015), 1692–1703.

[7] X. He and P. Niyogi. 2003. Locality Preserving Projections. In *Advances in Neural Information Processing Systems 16 (NIPS)*. 153–160.

[8] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen. 2011. Local Binary Patterns and Its Application to Facial Image Analysis: A Survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 41, 6 (2011), 765–781.

[9] P. Isola, D. Parikh, A. Torralba, and A. Oliva. 2011. Understanding the Intrinsic Memorability of Images. In *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2429–2437.

[10] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. 2014. What Makes a Photograph Memorable? *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 7 (July 2014), 1469–1482.

[11] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva. 2015. Understanding and Predicting Image Memorability at a Large Scale. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 2390–2398.

[12] H. Peng, K. Li, B. Li, H. Ling, W. Xiong, and W. Hu. 2015. Predicting Image Memorability by Multi-view Adaptive Regression. In *Proceedings of the 23rd ACM International Conference on Multimedia (MM '15)*. ACM, New York, NY, USA, 1147–1150.

[13] B. Scholkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. 2000. New Support Vector Algorithms. *Neural Comput.* 12, 5 (2000), 1207–1245.

[14] S. Shekhar, D. Singal, H. Singh, M. Kedia, and A. Shetty. 2017. Show and Recall: Learning What Makes Videos Memorable. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2730–2739.

[15] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. 4489–4497.