Mining Emotional Features of Movies

Yang Liu[1,2], Zhonglei Gu[3], Yu Zhang[4], Yan Liu[5]
[1]Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China
[2]Institute of Research and Continuing Education, Hong Kong Baptist University, Shenzhen, China
[3]AAOO Tech Limited, Hong Kong SAR, China
[4]Department of CSE, Hong Kong University of Science and Technology, Hong Kong SAR, China
[5]Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China
csygliu@comp.hkbu.edu.hk, allen.koo@aaoo-tech.com, zhangyu@cse.ust.hk
csyliu@comp.polyu.edu.hk

ABSTRACT

In this paper, we present the algorithm designed for mining emotional features of movies. The algorithm dubbed Arousal-Valence Discriminant Preserving Embedding (AV-DPE) is proposed to extract the intrinsic features embedded in movies that are essentially differentiating in both arousal and valence directions. After dimensionality reduction, we use the neural network and support vector regressor to make the final prediction. Experimental results show that the extracted features can capture most of the discriminant information in movie emotions.

1. INTRODUCTION

Affective multimedia content analysis aims to automatically recognize and analyze the emotions evoked by multimedia data such as images, music, and videos. It has a lot of real-world applications such as image search, movie recommendation, and music classification [3, 7–9, 11–14].

In this *2016 Emotional Impact of Movies Task*, the participants are required to design algorithms to predict the arousal and valence values of the given movies automatically. The dataset used in this task is the LIRIS-ACCEDE dataset (liris-accede.ec-lyon.fr). It contains videos from a set of 160 professionally made and amateur movies, shared under the Creative Commons licenses that allow redistribution [2]. More details of the task requirements as well as the dataset description can be found in [5, 10].

In this paper, we perform both global and continuous emotion predictions via a proposed supervised dimensionality reduction algorithm called Arousal-Valence Discriminant Preserving Embedding (AV-DPE), which learns the compact representations of the original data. After obtaining the low-dimensional features, we use the neural network and support vector regressor to predict the emotion values.

2. PROPOSED METHOD

In order to derive the intrinsic factors in movies that convey or evoke emotions along the arousal and valence dimensions, we propose a supervised feature extraction algorithm dubbed Arousal-Valence Discriminant Preserving Embedding (AV-DPE) to map the original high-dimensional representations into a low-dimensional feature subspace, in which the data with similar A-V values are close to each other, while the data with different A-V values are faraway from each other.

Let $\mathbf{x} \in \mathbb{R}^D$ be the high-dimensional feature vector of the movie, and $\mathbf{y} = [y^{(1)}, y^{(2)}]$ be the corresponding emotion label vector, where $y^{(1)}$ and $y^{(2)}$ denote the arousal value and valence value, respectively. Given the training set $\{(\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_n, \mathbf{y}_n)\}$, AV-DPE aims at learning a transformation matrix $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_d] \in \mathbb{R}^{D\times d}$ which is able to project the original D-dimensional data to an intrinsically low-dimensional subspace $\mathbb{Z} = \mathbb{R}^d$.

In order to describe the similarity between data samples, we define the following adjacency scatter matrix:

$$\mathbf{S}_a = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T, \quad (1)$$

where A_{ij} denotes the similarity between the i-th and j-th data points. In our formulation, we use the form of inner product between the corresponding label vectors associated with $\mathbf{x}_i$ and $\mathbf{x}_j$. To further normalize the similarity values into interval $[0, 1]$, we define the normalized adjacency matrix $\hat{\mathbf{A}}$ where

$$\hat{A}_{ij} = \langle \hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j \rangle = \langle \mathbf{y}_i/||\mathbf{y}_i||, \mathbf{y}_j/||\mathbf{y}_j|| \rangle. \quad (2)$$

The normalized adjacency scatter matrix is then defined as:

$$\hat{\mathbf{S}}_a = \sum_{i=1}^{n}\sum_{j=1}^{n} \hat{A}_{ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T. \quad (3)$$

Similarly, we define the normalized discriminant scatter matrix to characterize the dissimilarity between data points:

$$\hat{\mathbf{S}}_d = \sum_{i=1}^{n}\sum_{j=1}^{n} \hat{D}_{ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T, \quad (4)$$

where we simply define $\hat{D}_{ij} = 1 - \hat{A}_{ij}$.

In order to maximize the distance between data points with different labels while minimizing the distance between data points with similar labels, the objective function of AV-DPE is formulated as follows:

$$\mathbf{U} = \arg\max_{\mathbf{U}} \{tr((\mathbf{U}^T\hat{\mathbf{S}}_a\mathbf{U})^\dagger \mathbf{U}^T\hat{\mathbf{S}}_d\mathbf{U})\}, \quad (5)$$

where $tr(\cdot)$ denotes the matrix trace operation and $(\hat{\mathbf{S}}_a)^\dagger$ denotes the Moore-Penrose pseudoinverse of $\hat{\mathbf{S}}_a$ [6]. The optimization problem in Eq. (5) can be solved by some standard matrix decomposition techniques [6].

MediaEval 2016 Workshop, October 20-21, 2016, Hilversum, Netherlands.

Table 1: Results on global emotion prediction

Runs \ Criteria	Arousal		Valence	
	MSE	Pearson's CC	MSE	Pearson's CC
#1	1.18511707891	0.158772315634	0.235909661034	0.102487446458
#2	1.18260763366	0.174547894742	0.378511708782	0.378511708782
#3	1.46475414861	0.212414301359	0.267627565271	0.089311269390
#4	1.61515123698	0.201427253365	0.239352667040	0.133965496755

Table 2: Results on continuous emotion prediction

Runs \ Criteria	Arousal		Valence	
	MSE	Pearson's CC	MSE	Pearson's CC
#1	0.152869437388	0.0500544335696	0.125062204735	0.00901181966468
#2	0.128197164652	0.0557718765692	0.105905051008	0.0117374077757
#3	0.125552338276	0.0266523947466	0.139507683129	0.00139093558922
#4	0.293856466692	0.0266523946850	0.124565684871	0.0192993915142

3. EXPERIMENTS

In this section, we report the experimental settings and the evaluation results.

Global emotion prediction: we construct a 34-D feature set, including alpha, asymmetry_env, colorfulness, colorRawEnergy, colorStrength, compositionalBalance, cutLength, depthOfField, entropyComplexity, flatness, globalActivity, hueCount, lightning, maxSaliencyCount, medianLightness, minEnergy, nbFades, nbSceneCuts, nbWhiteFrames, saliencyDisparity, spatialEdgeDistributionArea, wtf_max2stdratio_{1-12} and zcr. Note that all above features are provided by the task organizers.

- Run #1: We use the original 34-D features as the input, and then use a function fitting neural network [1] with 100 nodes in the hidden layer for prediction. The Levenberg-Marquardt backpropagation function is used in training.
- Run #2: We use the original 34-D features as input, and then use the ν-support vector regression (ν-SVR) for prediction. In ν-SVR, the RBF kernel is utilized with the default setting from LIBSVM [4], i.e., $cost = 1$, $\nu = 0.5$, and γ is then set to be the reciprocal of the number of feature dimension.
- Run #3: We first use the proposed AV-DPE to reduce the original feature space to the 10-D subspace. Then utilize the neural network for prediction. The setting of neural network is the same as that in Run #1.
- Run #4: We first use the proposed AV-DPE to reduce the original feature space to the 10-D subspace. Then we use the ν-SVR for prediction. The setting of ν-SVR is the same as that in Run #2.

Continuous emotion prediction: we downsample the size of each video to 64×36. As a result, we have a 6912-D feature vector of RGB values for each frame.

- Run #1: We use the original 6912-D features as the input, and then use the neural network for prediction. The setting of neural network is the same as that in Run #1 of global emotion prediction.
- Run #2: We use the original 6912-D features as the input, and then use the ν-SVR for prediction. The setting of ν-SVR is the same as that in Run #2 of global emotion prediction.
- Run #3: We first use the proposed AV-DPE to reduce the original high-dimensional feature space to the 100-D subspace. Then we use the neural network for prediction. The setting of neural network is the same as that in Run #1 of global emotion prediction.
- Run #4: We first use the proposed AV-DPE to reduce the original high-dimensional feature space to the 100-D subspace. Then we use the ν-SVR for prediction. The setting of ν-SVR is the same as that in Run #2 of global emotion prediction.

Table 1 and Table 2 report the results of our system. From the tables we can see that after dimensionality reduction, the performance of the reduced features (Run #3 and Run #4) is generally worse than that of the original features (Run #1 and Run #2), which indicates that the emotion information in movies is relatively complex, and thus we may not be able to fully describe it using just a few dimensions. However, considering that the dimension of the reduced features is much less than that of the original features, we still can conclude that the learned subspace preserves rich discriminant information of the original feature space.

Moreover, from both tables we can observe that the neural network performs more robust than SVR after dimensionality reduction. The possible reason is that besides the discriminant ability, the neural network with the hidden layer has better representation ability of the original data than SVR, which is also of great importance in supervised learning tasks.

4. CONCLUSIONS

In this working notes paper, we have proposed a dimensionality reduction method to extract the emotional features from movies. By minimizing the distance between data points with similar emotion levels and maximizing the distance between data points with different emotion levels simultaneously, the learned subspace keeps most of the discriminant information and gives relatively robust results in both global and continuous emotion prediction tasks.

Acknowledgments

The authors would like to thank the reviewer for the helpful comments. This work was supported in part by the National Natural Science Foundation of China under Grant 61503317.

5. REFERENCES

[1] http://www.mathworks.com/help/nnet/ref/fitnet.html?requestedDomain=cn.mathworks.com.

[2] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, Jan 2015.

[3] L. Canini, S. Benini, and R. Leonardi. Affective recommendation of movies based on selected connotative features. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(4):636–647, April 2013.

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

[5] E. Dellandréa, L. Chen, Y. Baveye, M. Sjoberg, and C. Chamaret. The mediaeval 2016 emotional impact of movies task. In *Mediaeval 2016 Workshop*, 2016.

[6] G. H. Golub and C. F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[7] Y. Liu, Y. Liu, C. Wang, X. Wang, P. Zhou, G. Yu, and K. C. C. Chan. What strikes the strings of your heart? – multi-label dimensionality reduction for music emotion analysis via brain imaging. *IEEE Transactions on Autonomous Mental Development*, 7(3):176–188, Sept 2015.

[8] Y. Liu, Y. Liu, Y. Zhao, and K. A. Hua. What strikes the strings of your heart? – feature mining for music emotion analysis. *IEEE Transactions on Affective Computing*, 6(3):247–260, July 2015.

[9] R. R. Shah, Y. Yu, and R. Zimmermann. Advisor: Personalized video soundtrack recommendation by late fusion with heuristic rankings. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 607–616, 2014.

[10] M. Sjoberg, Y. Baveye, H. Wang, V. L. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, and L. Chen. The mediaeval 2015 affective impact of movies task. In *Mediaeval 2015 Workshop*, 2015.

[11] O. Sourina, Y. Liu, and M. K. Nguyen. Real-time eeg-based emotion recognition for music therapy. *Journal on Multimodal User Interfaces*, 5(1):27–35, 2012.

[12] X. Wang, J. Jia, J. Tang, B. Wu, L. Cai, and L. Xie. Modeling emotion influence in image social networks. *IEEE Transactions on Affective Computing*, 6(3):286–297, July 2015.

[13] K. Yadati, H. Katti, and M. Kankanhalli. Cavva: Computational affective video-in-video advertising. *IEEE Transactions on Multimedia*, 16(1):15–23, Jan 2014.

[14] S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian. Affective visualization and retrieval for music video. *IEEE Transactions on Multimedia*, 12(6):510–522, Oct 2010.