Analyzing Human Behavior in Subspace: Dimensionality Reduction + Classification

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

Automated detection of human behavior in a social setting has drawn considerable interests in recent years. In this working notes paper, we describe our system developed for human behavior analysis. The system is composed of two components: 1) a dimensionality reduction module that maps the original data to a subspace; and 2) a classifier module that classifies the test data based on the labels of training data in the learned subspace. The developed system is evaluated on the MediaEval 2018 Human Behavior Analysis Task.

1 INTRODUCTION

Automated detection of human behavior in a social setting has drawn considerable interests in recent years. Unlike the human behavior analysis focusing on a single person, detection of human behavior in a social setting emphasizes more on the dynamics between different participants in a social event, where indicators such as the participants' speech pattern, their body language, and movements of body can be used to deduce valuable implications in understanding how human behavior in a social setting can contribute to the personal and/or career progression of an individual. Naturally, analyzing audio content recorded during the social event would yield a series of valuable information, such as participants' speech pattern, the pitch, tone, and pacing of how each individual speak, or even content covered during the discussion, that would help in identifying potential social traits in an individual's personal and career development. However, these audio contents may very often contain sensitive information in which major security concerns may arise in recording and using such content. As a result, alternative measures are being explored to discover human behavior in social setting in a less privacy-invasive way. In the MediaEval 2018 Human Behavior Analysis Task [1], people's body movement, as recorded by a tri-axial accelerator, along with other accompanying visual features are provided to participants in a hope to derive effective alternative approaches to analyze human behavior in a social setting without the use of audio content.

2 APPROACH

In this section, we introduce our system designed for the human behavior analysis task. The developed system is composed of two components. The first component is a dimensionality reduction module that maps the original data to a subspace. The motivation of using dimensionality reduction to learn the subspace is that the original high-dimensional feature space often contains redundant or even noisy information, which may affect the efficiency and accuracy. In our system, we choose principal component analysis (PCA) [2, 3] for dimensionality reduction as it is efficient and easy to interpret. The second component a classifier module that classifies the test data based on the labels of training data in the learned subspace. In our system, we choose the nearest neighbor (NN) [4] method for classification because of, again, its efficiency and interpretability.

2.1 Dimensionality Reduction via Principal Component Analysis

Given the training data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$, where $\mathbf{x}_i \in \mathbb{R}^D$ denotes the feature vector of the i-th data sample, PCA aims to learn a $D \times d$ transformation matrix $\mathbf{W}$, which maps the original data to the d-dimensional subspace, with the data variance being maximumly preserved. To achieve this goal, PCA maximizes the following objective function:

$$\mathbf{W} = \arg\max_{\mathbf{W}} tr\Big(\mathbf{W}^T(\mathbf{X} - \bar{\mathbf{x}}\mathbf{1}_n^T)(\mathbf{X} - \bar{\mathbf{x}}\mathbf{1}_n^T)^T\mathbf{W}\Big), \qquad (1)$$

where $\bar{\mathbf{x}} = \sum_{i=1}^n \mathbf{x}_i/n$ and $\mathbf{1}_n$ denotes the $n \times 1$ vector with all entries being 1. By further introducing a scaling constraint $\mathbf{W}^T\mathbf{W} = \mathbf{I}$, the optimal $\mathbf{W}$ that maximizes Eq. (1) is composed of the normalized eigenvectors corresponding to the d largest eigenvalues of the following eigen-decomposition problem:

$$(\mathbf{X} - \bar{\mathbf{x}}\mathbf{1}_n^T)(\mathbf{X} - \bar{\mathbf{x}}\mathbf{1}_n^T)^T\mathbf{w} = \lambda\mathbf{w}. \qquad (2)$$

After obtaining the transformation matrix $\mathbf{W}$, we can map the original high-dimensional data sample $\mathbf{x}_i$ in both training and test sets to the low-dimensional subspace by: $\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i$.

2.2 Classification via Nearest Neighbor Method

For a given test data sample, NN assigns the class label of test sample's nearest neighbor in the training set to the test

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

sample. Specifically, given the low-dimensional representation of the training set, i.e., $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$, the label of a test data sample $\mathbf{y}$ is decided by the following function:

$$l(\mathbf{y}) = l\Big(\mathop{\arg\min}_{\substack{\mathbf{y}_i \\ i=1,\cdots,n}} d(\mathbf{y}, \mathbf{y}_i)\Big), \tag{3}$$

where $l(\mathbf{y})$ denotes the label of $\mathbf{y}$, and $d(\mathbf{y}, \mathbf{y}_i)$ denotes the distance between $\mathbf{y}$ and $\mathbf{y}_i$. In this paper, we utilize the widely used Euclidean distance as the distance metric.

3 RESULTS AND ANALYSIS

We evaluate the performance of our system on the MediaEval 2018 Human Behavior Analysis Task. The dataset is composed of two parts: 1) The development set with 54 subjects. The video for each subject is 22 minutes (i.e., $1,320$ seconds) long. So we have $54 \times 1,320 = 71,280$ training samples in total; 2) The test set with 16 subjects. The video for each subject is also 22 minutes (i.e., $1,320$ seconds) long. So we have $54 \times 1,320 = 21,120$ test samples in total.

We use three types of features to construct our original data representation: 1) Colorhist: we calculate the standard deviation of 20 frames' colorhist as the representative of that second, and the dimension is 128; 2) LBP: we calculate the standard deviation of 20 frames' LBP as the representative of that second, and the dimension is 256; 3) Accel: for each frame, this feature is 3-dimensional, and we concatenate these 3-D feature of all 20 frames as the representative of that second, the dimension is 60. For Acceleration, Video, and Fusion, we submit two runs for each of them.

- For Run 1 of Acceleration, we use 60-D Accel feature and perform NN classification directly.
- For Run 2 of Acceleration, we use PCA to project 60-D Accel feature to a 10-D subspace, and perform NN classification in the learned subspace.
- For Run 1 of Video, we use 384-D feature (Colorhist + LBP) and perform NN classification directly.
- For Run 2 of Video, we use PCA to project 384-D feature (Colorhist + LBP) to a 50-D subspace, and perform NN classification in the learned subspace.
- For Run 1 of Fusion, we use 444-D feature (Colorhist + LBP + Accel) and perform NN classification directly.
- For Run 2 of Fusion, we use PCA to project 444-D feature (Colorhist + LBP + Accel) to a 50-D subspace, and perform NN classification in the learned subspace.

Table 1 demonstrates the results (in terms of AUC) of our system on Run 1 and Run 2 of Accel, Video, and Fusion. The overall performance is far from satisfactory and has large room for improvement. There might be several reasons. First, the dimensionality reduction and classification methods used in our system are very simple and straightforward. They may have the advantages in implementation and interpretation. However, they may lack the ability to extract sufficient discriminative information from the original feature space for classification. Second, the label provided by NN is binary, whereas the evaluation criterion ROC-AUC requires probabilities. The inconsistency between them may further degrade the performance. Third, the feature set we have used may not be sufficient to capture all the discriminative information embedded in the original videos. In addition to the observation on overall performance, we also see that compared with the Video feature, the Accel feature plays a more important role in classification, even its dimension is low than the Video feature's dimension. Moreover, by comparing Run 1 and Run 2, we find that PCA does not really improve the performance, which motivates us to seek more powerful dimensionality reduction methods for the task in the future.

Table 1: Results (in terms of AUC) of Run 1 and Run 2 of Accel, Video, and Fusion on the MediaEval 2018 Human Behavior Analysis Task.

ID	Run 1			Run 2		
	Accel	Video	Fusion	Accel	Video	Fusion
2	0.5486	0.5300	0.5467	0.5502	0.5175	0.5426
3	0.5491	0.5306	0.5567	0.5197	0.5299	0.5515
15	0.5314	0.4998	0.5484	0.5434	0.5059	0.5541
17	0.5254	0.4912	0.5095	0.5396	0.5018	0.5210
26	0.5333	0.5078	0.5098	0.5249	0.5019	0.5109
39	0.5468	0.5483	0.5521	0.5542	0.5571	0.5400
40	0.5341	0.5152	0.5199	0.5546	0.5354	0.5379
43	0.5158	0.5027	0.5290	0.5296	0.5077	0.5376
51	0.5177	0.5540	0.5397	0.5045	0.5499	0.5403
54	0.5067	0.4967	0.5125	0.4750	0.4896	0.5346
59	0.5007	0.5342	0.5196	0.5293	0.5251	0.5125
65	0.5333	0.4975	0.5356	0.5318	0.5023	0.5360
67	0.5678	0.4858	0.5481	0.5412	0.4766	0.5494
80	0.5703	0.5111	0.5703	0.5558	0.4936	0.5735
83	0.5367	0.4859	0.5510	0.5164	0.4788	0.5631
85	0.5075	0.5017	0.5132	0.5188	0.5064	0.5125
Mean	0.5328	0.5120	0.5350	0.5305	0.5112	0.5380
Std	0.0198	0.0207	0.0186	0.0205	0.0225	0.0173

4 CONCLUSION

This working notes paper introduces our system design for identifying human behavior and shows the results of our system on the MediaEval 2018 Human Behavior Analysis Task. The unsatisfactory results motivate us to use more informative features and seek for more powerful dimensionality reduction and classification methods (such as deep neural networks) in the future.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61503317, in part by the General Research Fund (GRF) from the Research Grant Council (RGC) of Hong Kong SAR under Project HKBU12202417, and in part by the SZSTI Grant with the Projct Code JCYJ20170307161544087.

REFERENCES

[1] L. Cabrera-Quiros, E. Gedik, and H. Hung. 2018. No-Audio Multimodal Speech Detection in Crowded Social Settings task at MediaEval 2018. In *Mediaeval 2018 Workshop*.

[2] H. Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24, 7 (1933), 498–520.

[3] I. Jolliffe. 2002. *Principal component analysis*. Springer Verlag, New York.

[4] J. Laaksonen and E. Oja. 1996. Classification with learning k-nearest neighbors. In *Proceedings of International Conference on Neural Networks (ICNN'96)*, Vol. 3. 1480–1483.