Heterogeneous massive feature fusion on grassmannian manifold

Haichao Huang¹, Hongning Liu¹, Xiaoyun Kong², Xingdan Lou³ and Zepeng Wang⁴

¹State Grid Zhejiang Information & Telecommunication Company
²State Grid Zhejiang Electric Power Company
³Zhejiang Huayun Information Technology CO., LTD
⁴Department of CSIE, Hefei University of Technology

Abstract. Two issues remain unsolved on utilizing multimodal features for pattern recognition: the missing features and the curse of dimensionality. In this paper, we address the two issues by fusing the multimodal features on the Grassmann manifold. By defining grouping constraints on multimodal features, each multimodal feature vector is grouped into a set of subspaces, and is further represented as a point on the Grassmann manifold. To deal with missing features, L2-Hausdorff distance, a metric to compare multimodal feature vectors with different number of subspaces, is computed, and a kernel matrix can be obtained accordingly. Based on the kernel matrix, two feature selection criterions, one supervised and one unsupervised, are proposed to obtain a few representative features in the kernel space. Thus, the curse of dimensionality is alleviated. Experimental results on three multimodal dataset show the proposed feature fusion can outperforms the state-of-the-art by higher accuracy.

1. Introduction

When recognize an object, human usually adopts several types of features to identify the object more accurately and efficiently. For example, using color to discriminate oranges from bananas is difficult since both of them are yellow; similarly, it is difficult to discriminate apple from oranges using contour. If we fuse color and contour for classification, the three fruits will be recognized more easily and accurately. Motivated by this concept, researches in pattern recognition try to enhance the recognition accuracy by employing more than one single type of features, namely, multimodal features. In contrast with uni-modal feature, multimodal features contain richer information. Therefore, if we fuse multimodal features optimally, there is a great potential to achieve to a better recognition performance.

According to whether features in different modality will be reallocated, we classify the existing multimodal feature fusion methods into two categories, namely, multi-cue integration based fusion and modality identification based fusion, respectively.

Multi-cue based fusion integrates modalities by treating each type of feature as a modality, in the literature, a series of multi-cue based fusion methods have been proposed. In [1], each classifier’s accuracy in local regions of feature space is estimated, and then the most accurate local classifier is output as the final decision. In [2], a theoretical framework of fusing decisions from multiple classifiers is provided. Specifically, each fusing scheme can be explained as a special case of the framework. Furthermore, the sensitivity to estimation errors of different schemes is analyzed. By allocating each modality with a sub-classifier, score level feature fusion[3] quantize the multimodal feature vector into a set of class labels predicted by the sub-classifiers, and the final label is obtained by fusing the sub-classifiers’ labels through a combination strategy, such as the max rule or the majority vote rule. In
[4], each modality is represented as a graph and graphs from various modalities are integrated into a single graph by an adaptation scheme. A decision tree based modality integration method is proposed in [5]. To be specific, each modality is represented as a node in the decision tree and decisions at each node are taken using an accumulation scheme; decision from each node is multiplied by a coefficient that indicates the reliability of the corresponding modality. In [6], features from each modality is represented as one independent graph, and different learning tasks are formulated as inferring from the constants in every graph as well as supervision information. The above multi-cue based fusion method can alleviate the dimensionality curse. However, by allocating different types of features into different modalities, the features in each modality are obtained in a heuristic manner.

In this paper, we propose an approach of multimodal features fusion which can handle missing features in recognition and, at the same time, alleviate the curse of dimensionality. We try to deal with the multimodal features on the Grassmann manifold and fuse the multimodal features into kernel matrix. Targeting this aim, we group the multimodal features into a set of subspaces either by minimizing inter-group feature correlation and maximizing intra-group feature correlation or, a greedy method. Next, L2-Hausdorff distance, a metric to compare multimodal feature vectors with different number of subspaces, is computed, and a kernel matrix can be obtained accordingly. To handle missing features, if one subspace contains missing features, it is abandoned in distance computation. Finally, two feature selection criterion in kernel space, one for classification and the other for clustering, is proposed to obtain a few representative features in the kernel space. Thus, the curse of dimensionality can be alleviated.

1.1. Minimize the Inter-Group Feature Correlation and Maximize the Intra-Group Feature Correlation

When employing the multimodal features for clustering, due to the absence of class labels, the grouping scheme is developed by making use of the features correlation.

To quantize the correlation between multimodal features, two measures, one based on classic linear correlation and the other based on information theory [7], can be adopted. For the linear correlation, the most well-known measure is linear correlation coefficient. Given a pair of features \( X \) and \( Y \), the linear correlation coefficient \( r \) is given by the formula:

\[
r = \frac{\Sigma (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma (x_i - \overline{x})^2} \sqrt{\Sigma (y_i - \overline{y})^2}}
\]

where \( \overline{x} \) is the means of \( X \) and \( \overline{y} \) is the means of \( Y \). It is noticeable that the value of \( r \) is restricts between -1 and 1. If \( X \) and \( Y \) are completely correlated, then \( r \) take the value of 1 or -1; if \( X \) and \( Y \) are totally independent, then \( r \) take the value of 0.

The linear correlation is an optimal choice to represent features’ correlation, under the assumption that a pair of multimodal features are linear separable. However, in the real world, it is not safe to always assume linear correlation between features, since linear correlation measures may not be able to capture correlations that are not linear in nature. To overcome this shortcoming, an information-theoretical concept based correlation, which can be deemed as a measure of the uncertainty of a multimodal feature, is presented in our approach.

Given a multimodal feature \( X \), its entropy is defined as:

\[
EN(X) = -\Sigma_{n=1} p(x^n)\log_2(p(x^n))
\]

The conditional entropy of feature \( X \) is defined as:

\[
EN(X|Y) = -\Sigma_{n=1} p(y^n) \Sigma_{n=1} p(x^n|y^n)\log_2(p(x^n|y^n))
\]

We can represent the feature correlation by computing the information gain [8]:

\[
IG(X|Y) = EN(X) - EN(X|Y)
\]

And symmetrical uncertainty [1] can be applied to normalize IG:

\[
SU(X, Y) = \frac{IG(X|Y)}{EN(X)+EN(Y)}
\]
It is noticeable that these entropy-based measures require nominal features, but they can be applied to measure correlations between continuous features as well, if the values are discretized in advance. By defining symmetrical uncertainty, given D modalities $M_1, M_2, ..., M_D$, each containing a number of multimodal features, the inter-group feature correlation between a pair of modalities is defined as:

$$ C(M_i, M_j) = \delta \sum_{X_i \in M_i} \sum_{X_j \in M_j} SU(X_i, X_j) $$

(6)

where $X_i$ and $X_j$ are features belonging to modalities $M_i$ and $M_j$ respectively, and $\delta$ is a factor to normalize $C(M_i, M_j)$ between -1 to 1. The intra-group feature correlation within modality $M_i$ is defined as:

$$ C(M_i) = C(M_i, M_i) $$

(7)

Based on the definition of symmetrical uncertainty, a criteria to allocate the multimodal features into $|D|$ groups is proposed, to be specific, the inter-group feature correlation is minimized and the intra-group feature correlation is maximized. The criteria can be formulated into the following objective function:

$$ \min \sum_{i=1, j>i} |D| \left[ \frac{C(M_i, M_j)}{C(M_i)} + \frac{C(M_i, M_j)}{C(M_j)} \right] $$

(8)

It is noticeable that the objective function yields $|D|$ modalities, with minimal inter-modality correlation and balanced features in each modality. Let $n$ denotes the number of multimodal features, the computational complexity is $O(n^2)$, which is computational efficient.

2. Map Subspaces into Grassmann Manifold

2.1. Principle Angles on the Grassmann Manifold

As shown in Figure 1, the Grassmann manifold $G(m, D)$ is the set of $m$-dimensional linear subspaces of the $R^D$. The $m$-dimensional linear subspaces correspond to a point on Grassmann manifold. The point can be seen as a matrix with size $m \times D$.

Let $X_1, X_2$ be two matrices of size $D$ by $m$, there are $m$ principle angles. The $i$th principle angle can be defined as:

$$ \cos \theta_i = \max_{u_k \in \text{orth}(X_1)} \max_{v_k \in \text{orth}(X_2)} |u_k^T v_k|, \quad (i=1, ..., k) $$

(9)

where $\text{orth}(X)$ represents the orthogonal basis of $X$.

![Figure 1. Principle angles in Grassmann manifold.](image)

2.2. Distance in Grassmann Manifold

A real-valued function $d: X \times X \rightarrow \mathbb{R}$ is called a metric if:

1. $d(x_1, x_2) \geq 0$,
2. $d(x_1, x_2) = 0$, if and only if $x_1 = x_2$
3. $d(x_1, x_2) = d(x_2, x_1)$,
4. \( d(x_1, x_2) + d(x_2, x_3) \geq d(x_1, x_3) \),

for all \( x_1, x_2, x_3 \in X \).

Based on the definition of principle angle, a series of metrics are derived accordingly. In this paper, three distances are used, they are Projection distance[10], Binet-Cauchy distance[11] and L2-Hausdorff distance[12], respectively.

If \( x_1 \) and \( x_2 \) contain the same number of subspaces, then the distance can be computed by:

1. The projection distance is defined as:
   \[
   d_p(x_1, x_2) = \left( \sum_{i=1}^{m} \sin^2 \theta_i \right)^{1/2} = \left( m - \sum_{i=1}^{m} \cos^2 \theta_i \right)^{1/2}
   \] (10)

2. The Binet-Cauchy distance is defined as:
   \[
   d_{bc}(x_1, x_2) = (1 - \sum_{i=1}^{m} \cos^2 \theta_i)^{1/2}
   \] (11)

If \( x_1 \) and \( x_2 \) contain the different number of subspaces, then the distance can be computed by:

3. The L2-Hausdorff distance is defined as:
   \[
   d_{l2}(x_1, x_2) = \sqrt{\max(m, n) - \sum_{i=1}^{m} \sum_{j=1}^{n} (u_i^T v_j)^2}
   \] (12)

where \( u_1, u_2, ..., u_m \) and \( v_1, v_2, ..., v_n \) are orthonormal bases of \( U \) and \( V \) respectively. It is noticeable that the L2-Hausdorff distance is independent on the choice of basis.

By defining the distances on the Grassmann manifold, a kernel matrix can be constructed accordingly. To be specific, let \( n_{tr} \) and \( n_{te} \) be the number of training and testing samples respectively, the training kernel matrix is a \( n_{tr} \times n_{tr} \) matrix \( M_{tr} \); the testing kernel matrix is a \( n_{te} \times n_{tr} \) matrix \( M_{te} \). The kernel value of \( M_{tr} \) is computed as:

\[
M_{tr}(i, j) = \exp(-\lambda * d(i, j))
\] (13)

where \( \lambda \) is a parameter to be tuned, and \( d(i, j) \) is the distance between the \( i \)th and the \( j \)th training sample.

Similarly, the kernel value of \( M_{te} \) is computed as:

\[
M_{te}(k, l) = \exp(-\lambda * d(k, l))
\] (14)

where \( d(k, l) \) is the distance between the \( k \)th testing sample and the \( l \)th training sample.

3. Feature Selection in Kernel Space

By constructing kernel matrix on the Grassmann manifold, we can select a few representative features in kernel space. In this section, two feature selection methods in kernel space, one supervised and one unsupervised are proposed.

4. Experimental Results and Analysis

To validate the performance of the proposed Grassmann manifold based feature fusion, we carry out three multimodal recognition tasks: one speech emotion recognition and two objects recognitions in images. All the three experiments run on a system equipped with Intel E8500 and 4GB main memory. In addition, the algorithm of our approach is implemented on Matlab platform.

4.1. Multimodal Speech Emotion Classification

In this experiment, we collected 511 sentences speech. The dataset contains 511 sentences emotional speech recorded by a female mandarin speaker. 41 acoustic features [11] from 6 types: pitch, log energy, the first three formant frequency, MFCC, PLCC and LSFs are extracted. We select half of the sentences for training and leave the rest for testing. The recorded speech data are labeled carefully with five basic emotions: Anger, Happiness, Neutral, Sadness and Surprise. Speech preprocessing including pre-emphasizing, blocking and Hamming windowing are performed on each speech frame of 25ms with 5ms frame shift. We extract the multimodal acoustic features on each frame. The C4.5 classifier [5] is used to evaluate the selected features’ discriminative ability. We compare the kernel matrix of our algorithm with three representative kernel matrix (linear kernel, polynomial kernel and Gaussian kernel), which are learned from feature level fusions [2]. The hyper-parameters corresponding to the three
kernels were learned by cross-validation. For simplicity, the number of subspaces is set to 5 in this experiment.

Table 1. Comparison of recognition rate (percentage), Hap. Means Happiness, Neur. Neutral, Sad. Sadness, Sur. Surprise, Ave. Average.

|       | Anger | Hap. | Neur. | Sad | Sur. | Ave. |
|-------|-------|------|-------|-----|------|------|
| Full  | 0.901 | 0.766| 0.301 | 0.715 | 0.502 | 0.637 |
| ± 0.051 | ± 0.013 | ± 0.033 | ± 0.025 | ± 0.068 | ± 0.038 |
| Linear kernel | 0.344 | 0.232 | 0.188 | 0.311 | 0.165 | 0.248 |
| ± 0.012 | ± 0.008 | ± 0.016 | ± 0.013 | ± 0.021 | ± 0.014 |
| Poly. kernel | 0.357 | 0.246 | 0.212 | 0.501 | 0.319 | 0.327 |
| ± 0.024 | ± 0.018 | ± 0.015 | ± 0.019 | ± 0.034 | ± 0.022 |
| Gaus. kernel | 0.404 | 0.344 | 0.237 | 0.522 | 0.433 | 0.388 |
| ± 0.026 | ± 0.022 | ± 0.028 | ± 0.021 | ± 0.023 | ± 0.024 |
| Grass. kernel | 0.812 | 0.701 | 0.622 | 0.922 | 0.723 | 0.756 |
| ± 0.009 | ± 0.013 | ± 0.011 | ± 0.007 | ± 0.015 | ± 0.011 |

The experimental results in Table 1 is measured in error rate by averaging five repeats on this dataset. As Table 1 shows, the best average recognition accuracy was achieved by our approach. Recognition accuracy of our approach is competitive to the other methods on average. Furthermore, when comparing recognition accuracy of each emotion, our approach has partial advantages, such as neutral, sad and surprise. It is noticeable that features selected using the other three kernels are less discriminative than the features selected by our approach, since an obviously advantage in recognition accuracy is observed in our approach.

Figure 2. Comparison of the ability to handle missing features. Miss_i means the i-th subspace is missing.

As shown in Figure 2, we evaluate the testing sentences' recognition accuracy when features in one subspace are missing. Firstly, the recognition accuracy of the full feature set is the highest, since no feature is missing. Then, as we observed, if one subspace is removed, recognition accuracy decreases, and recognition accuracy of our approach is the highest of all. That is to say, comparing with the the other three kernels, our approach can best reduce the decrease of recognition accuracy from missing features.

5. Conclusion

In this paper, we propose an approach to address two issues in multimodal feature fusion, the missing features and the curse of dimensionality. First, the multimodal feature vector is grouped into a set of subspaces, which is represented as a point on the Grassmann manifold. Then, to deal with missing
features, L2-Hausdorff distance, is used to compare multimodal feature vectors with different number of subspaces. Next, to alleviate the curse of dimensionality, two feature selection algorithms, one supervised and one unsupervised, are proposed to obtain a few representative features. To validate our approach, experiments on three multimodal datasets is carried out. Experimental results show the proposed feature fusion can outperforms the state-of-the-art by higher accuracy.

6. References

[1] K. Woods, W. Kegelmeyer Jr, K. Bowyer, Combination of multiple classifiers using local accuracy estimates, newblock IEEE T-PAMI, 19(4), pages: 405–410, 1997.
[2] J. Kittler, M. Hatef, R. Duin, J. Matas, on combining classifiers, IEEE T-PAMI, 20(3), pages: 226–239, 1998.
[3] D. Greene, P. Cunningham, A matrix factorization approach for integrating multiple data views, Machine Learning and Knowledge Discovery in Databases, pages: 423–438, 2009.
[4] Z. Zhao, H. Liu, Multi-source feature selection via geometry dependent covariance analysis, JMRL, pages: 36–47, 2009.
[5] F. Bach, G. Lanckriet, M. Jordan, ‘Multiple kernel learning, conic duality, and the smo algorithm, in Proc. of ICML, p. 6. ACM, 2004.
[6] P. Gehler, S. Nowozin, on feature combination for multiclass object classification, in Proc. of ICCV, pages: 221–228, 2009.
[7] T. Xia, D. Tao, T. Mei, Y. Zhang, Multiview spectral embedding, IEEE T-SMC-B, 40(6), pages 1438–1446, 2010.
[8] B. Xie, Y. Mu, D. Tao, K. Huang, m-sne: Multiview stochastic neighbor embedding, IEEE T-SMC-B, 41(4), pages: 1088–1096, 2011.
[9] X. Zhou, B. Bhanu, Feature fusion of side face and gait for videobased human identification, Pattern Recognition, 41(3), pages: 778–795, 2008.
[10] W. Zuo, H. Zhang, D. Zhang, K. Wang, and Post-processed lda for face and palmprint recognition: What is the rationale, Signal Processing, 90(8), pages: 2344–2352, 2010?
[11] Y. Wu, E. Chang, K. Chang, J. Smith, Optimal multimodal fusion for multimedia data analysis, ACM Multimedia, pages: 572–579, 2004.
[12] N. Cristianini, B. Scholkopf, Support vector machines and kernel methods: the new generation of learning machines, Ai Magazine, 23(3), pages 31–41, 2002.