Multimodal Feature Fusion for 3D Shape Recognition and Retrieval

Shuhui Bu, Shaoguang Cheng, Zhenbao Liu, and Junwei Han
Northwestern Polytechnical University, China

A 3D feature-learning framework combines the different modality data that characterizes a 3D shape to promote the discriminability of unimodal features.

Three-dimensional shapes are used extensively in fields such as mechanical design, multimedia games, architecture, and medical diagnosis.[1] All of these applications need to store, recognize, and retrieve 3D models effectively and automatically. Because the characteristics of 3D shapes differ from those of text and images, traditional classification and retrieval techniques cannot be applied directly to 3D objects. Hence, 3D shape analysis remains a challenging issue.

Researchers have proposed numerous solutions to 3D shape recognition, matching, and retrieval problems.[1,2] Although tremendous advancements have been made, current methods are still far from satisfactory for applying 3D objects in more realms. Geometry- and view-based methods, for example, only use partial information from a 3D object (see the "Geometry- and View-Based Methods for 3D Shape Analysis" sidebar). Geometry-based methods use the complex topological structure and geometric properties of the 3D model but ignore the visual similarities between 3D objects. Conversely, view-based methods only consider the visual characteristics of a model from different viewing angles. These methods neglect either the extrinsic features or intrinsic properties of 3D objects. What will happen if we combine different modality information in a creative and effective way?

Information in the real world has various manifestation modalities. Each of these typically carries different information and is rarely independent of others. For example, video contains visual and audio signals, images are often associated with captions and tags, and 3D models can be described by multiview images captured from different angles and 3D shape features. Because these totally different modalities depict the same object, some highly nonlinear relationships exist between them. However, different modalities have different representations and structures. For example, images are often represented with real-valued pixel intensities or the outputs of feature detectors, whereas 3D shapes are usually represented with 3D features that contain information about geometric attributes and topological structures. This makes it hard to discover the nonlinear relationships between features across modalities.

This article proposes fusing the different modality data of 3D shapes into a deep learning framework. Our core idea is to better mine the deep correlations of different modalities. High-level features are first extracted using two deep belief networks (DBNs), one for geometry-based modality with the input of a geodesic-aware bag of features (GA-BoF) and the other for view-based modality with the input of a bag of visual feature (BoVF).[3] We then use a restricted Boltzmann machine (RBM)[4] to associate the high-level features from each modality. Our method fuses intrinsic and extrinsic features to provide complementary information so better discriminability can be reached. Results from experiments on 3D shape retrieval and recognition tasks indicate that the proposed method can improve performance.

Multimodal Feature Extraction and Fusion

The proposed multimodal feature extraction and fusion method is a three-step process. Figure 1 illustrates the approach's pipeline.

Geometry-Based Feature Generation

We adopt a scale-invariant heat kernel signature and average geodesic distance as the low-level 3D shape descriptors to generate middle-level features. These two local descriptors are

1070-986X/14/$31.00 © 2014 IEEE

Published by the IEEE Computer Society

Geometry- and View-Based Methods for 3D Shape Analysis

Geometry- and view-based methods are popular solutions to 3D shape recognition, matching, and retrieval problems.[1,2].

Geometry-based analysis methods usually extract local or global descriptors and then train classifiers using these descriptors or calculate descriptor similarity for shape classification and retrieval.[3,4] These methods require high-quality descriptors, which influence the performance dramatically. Thus, the crucial problem in geometry-based approaches is how to define sensitive, unique, stable, and efficient shape descriptors that are robust against isometric transformation.

Instead of using the properties of the 3D model itself, view-based methods assume that if two 3D models are geometrically similar, they will also look similar from corresponding angles.[5,6] Biao Leng and his colleagues use deep belief networks (DBNs) to improve the performance of view features in 3D shapes.[7] This approach doesn't require geometric attributes or topological relationships, so it can handle 3D models with degeneration, holes, and missing patches. Generating highly discriminative descriptors for 3D objects typically requires capturing a large number of views, and consequently we need an effective way to organize and discover relationships between these views. With the rapid progress in machine learning, feature learning for local and global descriptors, which can improve the discriminability of the original descriptors, is becoming a hot research topic.[3–9]

References

1. J.W. Tangelder and R.C. Veltkamp, "A Survey of Content Based 3D Shape Retrieval Methods," *Multimedia Tools and Applications*, vol. 39, no. 3, 2008, pp. 441–471.
2. Z. Liu et al., "A Survey on Partial Retrieval of 3D Shapes," *J. Computer Science and Technology*, vol. 28, no. 5, 2013, pp. 836–851.
3. A.M. Bronstein et al., "Shape Google: Geometric Words and Expressions for Invariant Shape Retrieval," *ACM Trans. Graphics* (TOG), vol. 30, no. 1, 2011, article no. 1.
4. U. Castellani et al., "Sparse Points Matching by Combining 3D Mesh Saliency with Statistical Descriptors," *Computer Graphics Forum*, vol. 27, no. 2, 2008, pp. 643–652.
5. H. Laga, "Semantics-Driven Approach for Automatic Selection of Best Views of 3D Shapes," *Proc. 3rd Eurographics Conf. 3D Object Retrieval*, 2010, pp. 15–22.
6. V. Barra and S. Biasotti, "Learning Kernels on Extended REEB Graphs for 3D Shape Classification and Retrieval," *Proc. Eurographics Workshop 3D Object Retrieval*, 2013, pp. 25–32.
7. H. Laga, M. Mortara, and M. Spagnuolo, "Geometry and Context for Semantic Correspondences and Functionality Recognition in Man-made 3D Shapes," *ACM Trans. Graphics* (TOG), vol. 32, no. 5, 2013, article no. 150.
8. S. Bu et al., "Shift-Invariant Ring Feature for 3D Shape," *Visual Computer*, vol. 30, no. 6–8, 2014, pp. 867–876.
9. R. Litman and A.M. Bronstein, "Learning Spectral Descriptors for Deformable Shape Correspondence," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, 2014, pp. 171–180.

robust against nonrigid and complex shape deformations. More importantly, they both consider the global shape information, which is necessary for global shape retrieval.

Scale-Invariant Heat Kernel Signature. The heat kernel signature (HKS) is derived from a heat diffusion equation using a Laplace-Beltrami operator on surfaces, which has the advantages of providing rich local geometric information, invariance to isometric deformation, and multiscale characteristics.[5] However, the HKS is sensitive to the shape's scale. To cope with this problem, Michael Bronstein and Iasonas Kokkinos proposed a scale-invariant heat kernel signature (SI-HKS) by Fourier transform of the difference of the HKS.[6]

Average Geodesic Distance. The average geodesic distance (AGD) is initially introduced for the purpose of shape matching. However, the AGD is not robust when using extremum as a normalization factor; for example, using the intraclass geometric variations make the local descriptor change easily. It is therefore difficult to apply the AGD to generate a bag of words (BoW) from a set of models. We modify the normalization factor to the mean of geodesic distances between all pairs of vertices to cope with this limitation. For any model, the modified AGD descriptor has a fixed mean value 1.

Low-Level Descriptors. We concatenate the first six frequency components of the SI-HKS and AGD descriptors to form a low-level shape descriptor as

$$F(x_i) = (\mathrm{SIHKS}(x_i)[\omega_1, \ldots, \omega_6]), \mathrm{AGD}(x_i) \quad (1)$$

where the feature dimension is $M = 7$.

Geodesics-Aware Bag of Features. In the next step, we compute a BoF to represent the occurrence probability of geometric words, and adopt *k*-means to generate them. After obtaining the geometric words $C = \{c_1, c_2, \ldots, c_K\}$ of size K, we quantize the low-level descriptor space to obtain a compact representation. For

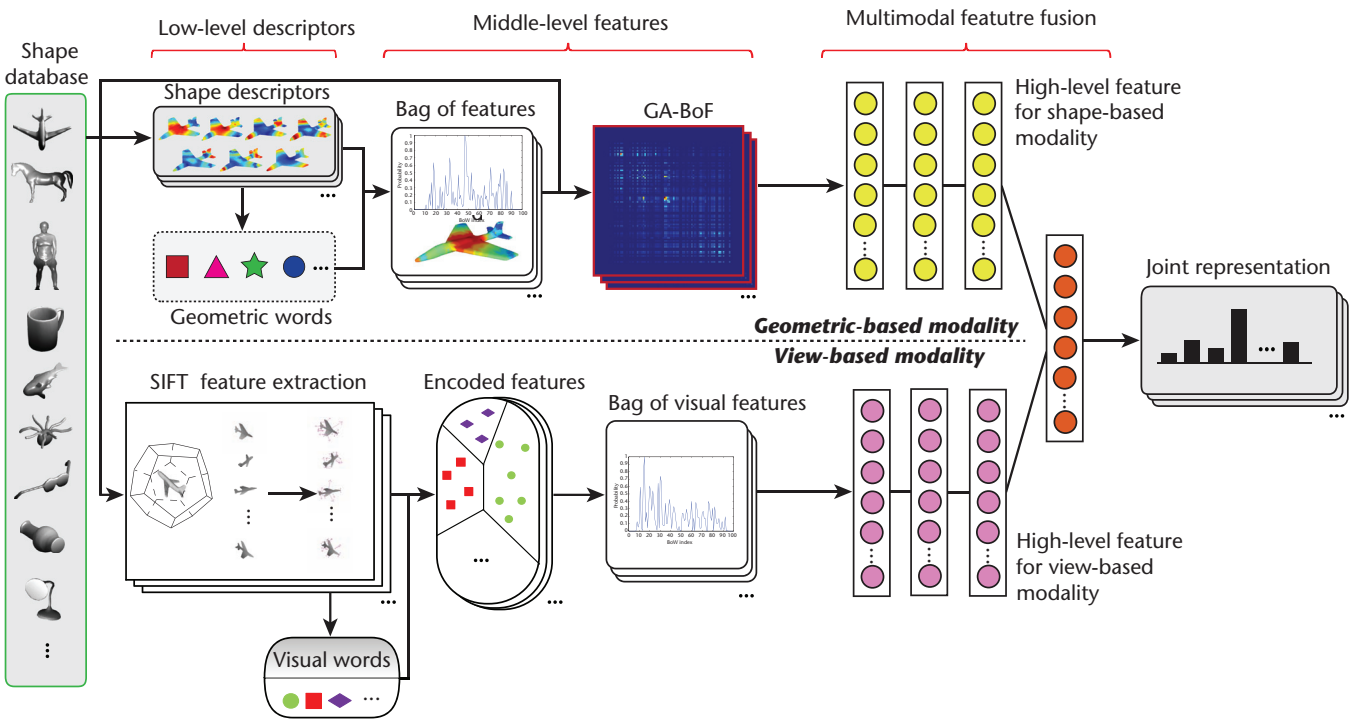

Figure 1. Flowchart of the proposed method. Only the offline training process is illustrated.

each point $x \in X$ with the descriptor $F(x)$, we define the feature distribution $\phi(x) = (\phi_1(x), \ldots, \phi_K(X))^T$ as a $K \times 1$ vector whose entries are

$$\varphi_i(x) = c(x) \exp\left(-\frac{\|F(x) - c_i\|_2^2}{k_{\mathrm{BoF}}\delta_{\min}^2}\right) \quad (2)$$

where the constant $c(x)$ is selected to satisfy $\|\phi(x)\|_1 = 1$.

The disadvantage of using BoF is it only considers the occurrence distribution of words and ignore the structural relationship between them, thus decreasing their discrimination. For geometric shapes, only features of the vertex are used, which limits their descriptive capability. Inspired by Shape Google,[7] we use the geodesics on the mesh to measure the spatial relationship between each pair of BoFs on the vertices. Unlike Shape google, we consider geodesic distance instead of heat kernel to avoid any influence of time scale and shape size, and we use the GA-BoF:

$$v(X) = N(X) \sum_{x_i \in X} \sum_{x_j \in X} \varphi(x_i)\varphi(x_j) \exp\left(-k_{gd}\frac{g_d(x_i, x_j)}{\sigma_{gd}}\right) \quad (3)$$

where $N(X)$ is a normalization factor that assigns features a fixed maximum value of 1; σ_{gd} is the maximal geodesic distance of any pair of vertices on the mesh; and k_{gd} denotes the decay rate of distances, which is selected empirically.

The resulting v is a $K \times K$ matrix, which represents the frequency of geometric words i and j appearing within a specified geodesic distance. This expression provides occurrence probability of geometric words and the relationship between them. Moreover, it provides a position-independent representation of shape, in which the positional independence denotes that the middle-level feature is irrelevant to the order of low-level features or vertices.

View-Based Feature Extraction

The view-based feature extraction for 3D models in our algorithm follows several steps.

Shape Preprocessing. For a given 3D mesh, we translate the center of its mass to the origin and then scale the maximum polar distance of the points on its surface to one. We don't perform rotation normalization but compensate for this to some extent, as we describe next.

Image Rendering from Multiple Views. Depth images are rendered from 20 vertices of a regular dodecahedron with mass center that is also located in the origin. To make the feature robust against rotation, we rotate the regular

dodecahedron 10 times and extract views in each position. The rotation angle must be set carefully to ensure that all the cameras are distributed uniformly and cover different viewing angles for the 3D model. We use a strategy similar to the light field descriptor (LFD) to extract views,[8] but unlike LFD, we discard the binary images and only use the depth images. Hence, a 3D object is represented by 200 depth images of size 256×256.

SIFT Feature Extraction. After rendering depth images, we extract the scale and rotation invariant visual features using the scale invariant feature transform (SIFT) algorithm. We set all the parameters of the SIFT algorithm to default, which produces a feature vector with 128 elements. The SIFT descriptor is robust against image noise and illumination changes. Moreover, it is stable under various changes of viewing angles,[9] which can compensate for the lack of rotation normalization. In our experiments, a 3D model is approximately represented with 5,000 to 7,000 SIFT features.

Bag-of-Visual-Features Generation. We code each SIFT feature extracted from the 3D models as a visual word by searching for the nearest neighbor in the vocabulary. Prior to encoding features, the vocabulary is learned with a *k*-means algorithm by clustering the features collected from every view of every model into a specified number of words. To obtain satisfactory discrimination, we set the number of words from 1,000 to 3,000 in our experiments. In addition, we randomly chose 50 percent of the models in each category to generate the visual vocabulary because of the tremendous computing workload of the clustering process.

After feature encoding, we collect the frequencies of words generated from a model in a histogram, which becomes a feature vector (the BoVF) for the 3D model.

Multimodal Feature Data Fusion

The direct way to train a multimodal model for 3D models is to build an RBM or DBN over the concatenated geometry- and view-based features. Because a joint model trained this way is limited as a shallow model, it is too hard to represent the highly nonlinear correlations and different statistical properties between both modalities. In our work, to associate geometry- and view-based data comprehensively, we first extract high-level features from shallow-level descriptors for each modality. As a result, information from a specific modality is weakened and more information in high-level features reflects the attributes of the 3D models. In other words, high-level features remove the modality-specific information and reserve only the attributes of the 3D models.

Unlike traditional deep learning methods, we don't use raw data as input for the deep learning structure. For view-based modality, our strategy for generating view images means that the order of raw images captured from views is not directly comparable. Every 3D model has many postures in the space. The images captured from the same angle of two similar shapes might differ significantly. Hence, it is not suitable to learn high-level features from raw image data. For geometry-based modality, because a 3D mesh has a graph structure, it is difficult to find a beginning position and make a comparable sequence. Therefore, it is difficult to learn high-level features from raw 3D shape data.

Because deep learning can extract deep structural information from features or raw data,[3,10] it is suitable for generating high-level features that can boost their discrimination ability. For geometry-based modality features, the GA-BoF can be regarded as a relationship matrix, with each entry representing the occurrence probability of two geometric words within a specified geodesic distance. Furthermore, all the shapes have the same size GA-BoFs, and this feature is invariant to the order and number of vertices on the mesh. Therefore, it is appropriate to construct a deep learning network. For view-based modality features, however, the BoVFs reflect the visual feature distribution of view images generated from each 3D shape and can be treated as input for deep learning.

The right side of Figure 1 shows the architecture of the suggested multimodal feature fusion. It contains two modality inputs: GA-BoFs and BoVFs. Each input is processed by a DBN. At the top of the DBNs, a RBM is used to learn the joint representation for the 3D model.

For each DBN, the bottom-layer RBM is trained with the input data, and the activation probabilities of hidden units are treated as input data for training the upper-layer RBM. The activation probabilities of the second-layer RBM are then used as the visible data input for the third-layer RBM and so on. After obtaining the optimal parameters for each DBN, the

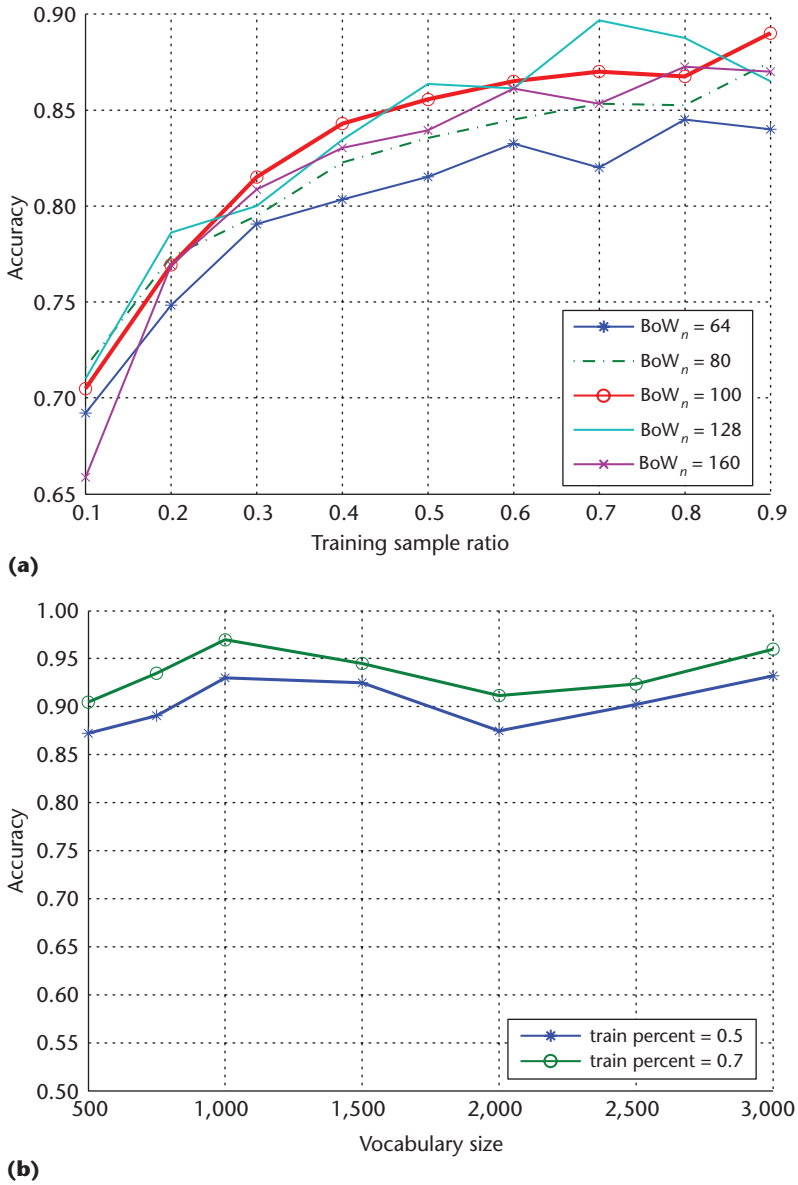

Figure 2. Average classification accuracy with different BoW_n on SHREC 2007. (a) Only geometry-based features, and (b) only view-based features.

newly input GA-BoF or BoVF is processed layer by layer until reaching the final layer. The last layer's output, $o(\mathbf{X}_{\text{shape}})$ and $o(\mathbf{X}_{\text{view}})$, is used as the high-level features for geometry- and view-based modality.

After these operations, an RBM is used to associate both modalities. First, high-level features for each modality are concatenated as $(o(\mathbf{X}_{\text{shape}}), o(\mathbf{X}_{\text{view}}))$. Next we input $(o(\mathbf{X}_{\text{shape}}), o(\mathbf{X}_{\text{view}}))$ into an RBM to learn a joint representation $o(\mathbf{X}_{\text{joint}})$. Because $o(\mathbf{X}_{\text{joint}})$ is generated from the geometry- and view-based modality features, it contains both intrinsic properties of the 3D model itself and extrinsic visual similarity and thus is more discriminative and robust.

For recognition tasks, we perform one-versus-all classification using Softmax regression on the learned joint representation. For the retrieval task, we use L_2, the distance of the joint representation, to measure the similarity of two shapes, **X** and **Y,** as

$$d_s(\mathbf{X}, \mathbf{Y}) = \left\| o(\mathbf{X}_{\text{joint}}) - o(\mathbf{Y}_{\text{joint}}) \right\|_2. \quad (4)$$

Experiments

To assess the proposed method, we use standard 3D shape benchmarks and a mixed dataset to evaluate classification and retrieval performance. The mixed dataset has more than 1,400 3D models, which consist of shapes from SHREC 2007, SHREC 2011, and the McGill database. The whole dataset is divided into 42 categories and each category contains approximately 10 to 30 meshes. We performed several experiments to select optimal parameters and then used the optimal parameters to evaluate the classification and retrieval performance.

To speed up the calculation, we implemented a deep learning toolbox, in which all matrix operations were carried out on the GPU using the Cudamat library. (The source code of our deep learning toolbox is available at *https://github.com/shaoguangcheng/DeepNet*.) All experiments were conducted on the platform with 8 Gbytes of memory and an Intel i3 core processor. For 400 shapes, the model learning took about 120 seconds, whereas the recognition and retrieval cost less than 1 second.

Optimal Parameters Selection

We first decided the optimal parameters for each modality using geometry- and view-based features individually to perform shape classification. We used average classification accuracy as the evaluation metric for the following experiments. The training data was randomly selected from the SHREC 2007 dataset, and the remaining data was treated as test data.

In the first experiment, we checked how the number of words affects performance. For the geometry-based modality, we set the number to 64, 80, 100, 128, and 160, respectively; Figure 2a shows the results. For the view-based modality, we set the number to 500, 750, 1,000, 1,500, 2,000, 2,500, and 3,000 separately; Figure 2b shows these results. As we can see, a small dictionary size generally leads to lower classification accuracy. Although a larger dictionary size

achieves better performance, the calculation time increases rapidly, resulting in low computation performance.

We also needed to select k_{BoF} and k_{gd} for the geometry-based feature extraction process. In the second experiment, we used different k_{BoF}, which controls how closely BoWs are selected as the BoF to evaluate the classification accuracy of training and testing data, using 100 BoWs. The results are plotted in Figure 3a. Next, we study the effects of using different k_{BoF}. This parameter indicates the decay rate for calculating the GA-BoF. Figure 3b shows the classification accuracy under different k_{gd}.

Based on the results of these experiments, we selected $\mathrm{BoW}_n = 100$, $k_{\mathrm{BoF}} = 2$, and $k_{gd} = 10$ as optimal parameters for geometry-based feature extraction, and $\mathrm{BoW}_n = 1,000$ for view-based feature generation. The following experiments were performed with these optimal parameters.

In our experiments, the number of DBN nodes and layers are set empirically. The geometry-based pathway consists of a Gaussian RBM with 5,050 visible units followed by two layers of 5,000 and 2,000 units. (Every 3D shape GA-BoF is a symmetric matrix, each entry of which represents the occurrence probability of two geometric words within a specified geodesic distance. So if we set $\mathrm{BoW}_n = 100$, the number of unique elements in GA-BoF is 5,050.) The view-based pathway also consists of a Gaussian RBM with 1,000 visible units followed by two layers of 1,000 and 800 units. The joint layer contains 2,800 hidden units.

Experiments on Classification

For the classification experiments, we randomly selected 50 percent of the models in each category as training samples and used the remaining models as test data. Table 1 lists the average classification accuracies of the proposed method and several other approaches on the SHREC 2007, SHREC 2011, and McGill datasets.

From Table 1, we can clearly conclude that the proposed multimodal feature fusion method achieves much better classification performance than using a single modality feature. This is because the geometry- and view-based modalities only reflect partial properties of the 3D model. We can obtain more discriminative power when both different modalities are considered. Moreover, we also use support vector machines (SVMs) to perform shape classification with the input of concatenated BoVFs and GA-BoFs. This experiment demonstrates that the suggested method is superior to the more common feature fusion approach. Among the three datasets, the results on SHREC 2011 have the best performance because the shapes only contain articulated deformation and the shape variance is small.

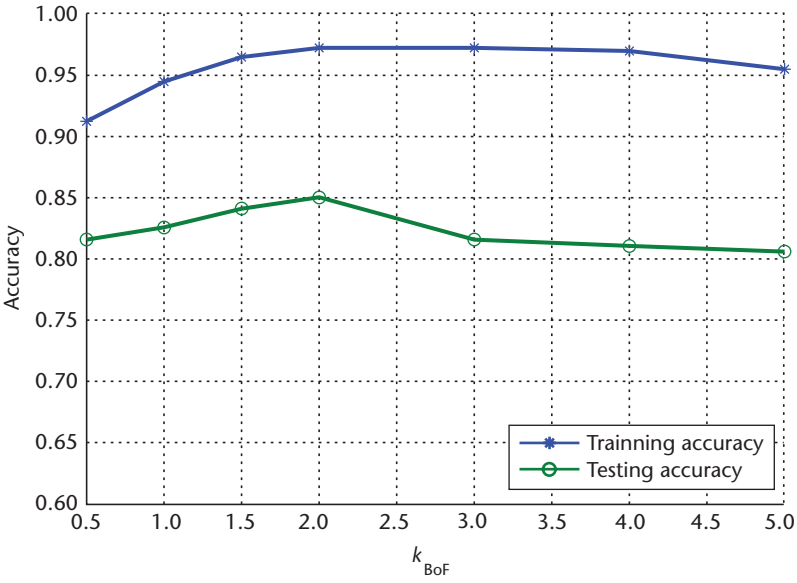

Figure 3. Average classification accuracy with different (a) k_{BoF} *and (b)* k_{gd} *on SHREC 2007.*

Experiments on Retrieval

For the retrieval task, we used the models trained in the classification experiments to calculate the joint representation for every 3D

Table 1. Average classification results of the proposed method and other approaches.

Method	SHREC 2007 (%)	SHREC 2011(%)	McGill (%)
Only geometry-based modality features	85.00	99.67	90.69
Only view-based modality features	93.00	97.00	89.00
Support vector machine (SVM) with multimodal features	80.00	92.17	78.51
Proposed method	**97.25**	**99.83**	**95.54**

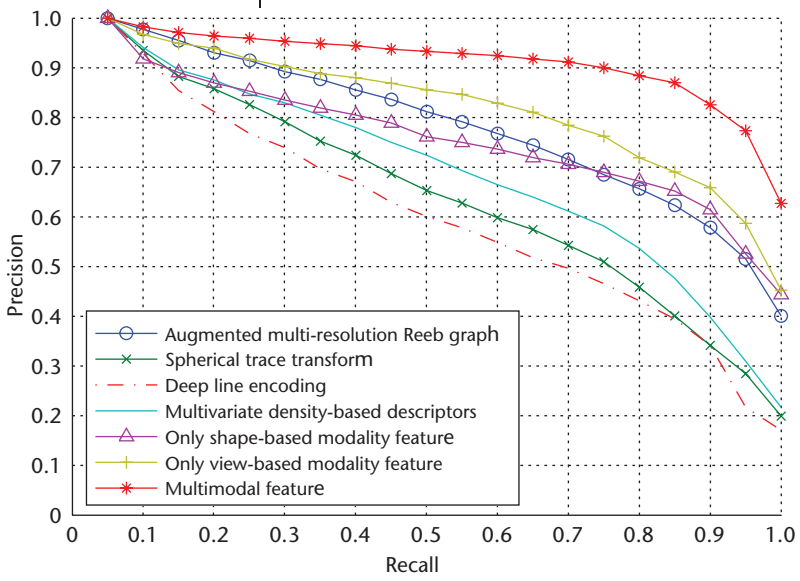

Figure 4. Recall-precision curves of some current methods and the proposed method on SHREC 2007. Using the joint representation increases the intraclass similarity while reducing the interclass similarity.

*Table 2. Retrieval performance of the proposed method using standard measures on SHREC 2007.**

Method	NN (%)	FT (%)	ST (%)	E (%)	DCG (%)
Only geometry-based modality features	83.75	66.81	39.92	56.34	89.91
Only view-based modality features	95.00	72.10	42.79	60.17	93.41
Proposed method	**97.50**	**83.29**	**46.28**	**66.54**	**96.73**

*Nearest neighbor (NN), first tier (FT), second tier (ST), E-measure (E), and discounted cumulative gain (DCG).

shape. We obtained the similarity between two models using Equation 4.

Evaluation Metrics. Six standard evaluation metrics are used to assess the performance of the recommended method. They are precision-recall curve, nearest neighbor (NN), first tier (FT), second tier (ST), E-measure (E), and discounted cumulative gain (DCG).

Experiments on SHREC 2007. First, we used the SHREC 2007 dataset to evaluate the proposed method's retrieval performance. Figure 4 plots the recall-precision curves of some state-of-the-art approaches and our method. From the figure, we can clearly see that the suggested method achieves the best retrieval results overall. If we use only a geometry-based modality feature or a view-based modality feature for the retrieval experiments, the performance shows no obvious improvement over its competitors. This is mainly because a unimodal feature can merely deliver specific information about a 3D shape. Because our recommended approach fuses geometry- and view-based modality information, the joint representation contains both intrinsic properties and extrinsic attributes of 3D models. Using the joint representation increases the intraclass similarity while reducing the interclass similarity, and consequently the retrieval performance is improved.

Table 2 lists the numerical evaluation measures. As the table shows, the measures are higher when we use multimodal features than when we use unimodal features. The average improvement of the DCG index is 5.07 percent, which demonstrates that the suggested method can improve retrieval performance by using multimodal features. We also find that nearest neighbor has the least improvement of all the evaluation indexes. This is mainly because the nearest neighbor only checks the validity of the nearest of the retrieval results, whereas our proposed method can ameliorate the whole retrieval performance.

Experiments on Mixed Dataset. Finally, we use a mixed 3D shape dataset as testing data to

evaluate the proposed approach's generalization ability and robustness. The mixed dataset consists of SHREC 2007, SHREC 2011, and McGill, thus is much more challenging. Figure 5 and Table 3 shows the results. Analyzing these measure indexes, we find the suggested approach has the best performance.

Conclusion

To learn the joint representation for 3D shapes, we propose a novel multimodal feature extraction and fusion method for 3D shapes. However, the information carried by each modality feature is not identical, as the experiments on the SHREC 2007 dataset demonstrate, where the view-based modality feature contains more information than geometry-based modality. Therefore, it is necessary to model the importance of different modalities. Moreover, in the proposed method, features for deep learning are global, so local information of 3D shapes is missing. Additionally, the optimal word number of BoVF and GA-BoF cannot be automatically decided.

Thus far, we have only investigated geometry- and view-based modalities in our framework. Although promising results were obtained, we could achieve better recognition and retrieval performance by adding other modality features, such as sketch-based features. This will be a subject of our future research. In addition, to better describe 3D shapes, we will explore the possibility of combining global and local features from each modality in our framework. **MM**

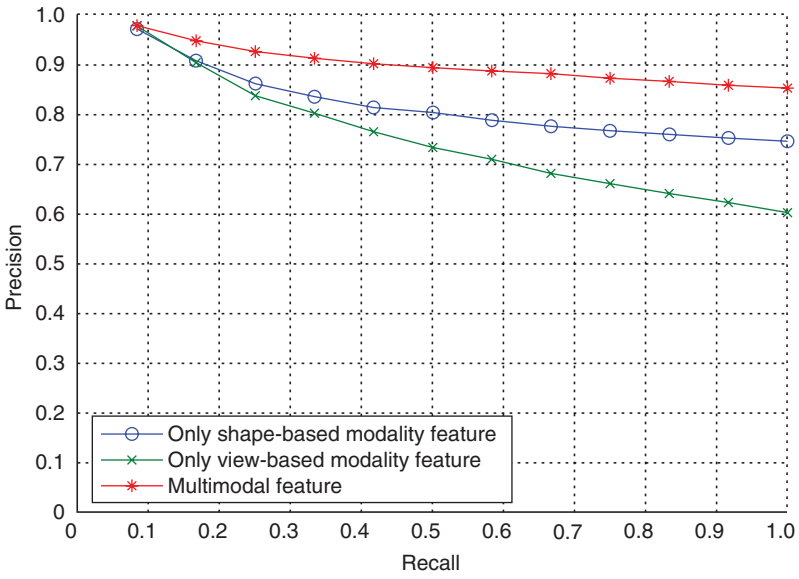

Figure 5. Recall-precision curves of proposed method on a mixed dataset. The mixed dataset consists of SHREC 2007, SHREC 2011, and McGill.

*Table 3. Retrieval performance of proposed method using standard measures on a mixed dataset.**

Method	NN (%)	FT (%)	ST (%)	E (%)	DCG (%)
Only geometry-based modality features	83.70	60.18	35.96	54.87	86.93
Only view-based modality features	82.66	45.06	29.67	43.27	81.61
Proposed method	**89.75**	**72.33**	**42.24**	**63.41**	**92.63**

*Nearest neighbor (NN), first tier (FT), second tier (ST), E-measure (E), and discounted cumulative gain (DCG).

Acknowledgments

This work is partly supported by grants from the National Natural Science Foundation of China (61202185, 61003137, 91120005, and 61473231), the Fundamental Research Funds for the Central Universities (310201401-(JCQ01009, JCQ01012)), Shaanxi Natural Science Fund (2012JQ8037), and the Open Project Program of the State Key Lab of CAD&CG (A1306) at Zhejiang University.

References

1. J.W. Tangelder and R.C. Veltkamp, "A Survey of Content Based 3D Shape Retrieval Methods," *Multimedia Tools and Applications*, vol. 39, no. 3, 2008, pp. 441–471.
2. Z. Liu et al., "A Survey on Partial Retrieval of 3D Shapes," *J. Computer Science and Technology*, vol. 28, no. 5, 2013, pp. 836–851.
3. G.E. Hinton and R.R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313, no. 5786, 2006, pp. 504–507.
4. G.E. Hinton, "Training Products of Experts by Minimizing Contrastive Divergence," *Neural Computation*, vol. 14, no. 8, 2002, pp. 1771–1800.
5. J. Sun, M. Ovsjanikov, and L. Guibas, "A Concise and Provably Informative Multi-scale Signature Based on Heat Diffusion," *Computer Graphics Forum*, vol. 28, no. 5, 2009, pp. 1383–1392.
6. M.M. Bronstein and I. Kokkinos, "Scale-Invariant Heat Kernel Signatures for Non-rigid Shape Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (CVPR), 2010, pp. 1704–1711.
7. A.M. Bronstein et al., "Shape Google: Geometric Words and Expressions for Invariant Shape Retrieval," *ACM Trans. Graphics* (TOG), vol. 30, no. 1, 2011, article no. 1.

8. Y.-T. Shen et al., "3D Model Search Engine Based on Lightfield Descriptors," *Proc. Eurographics Interactive Demos*, 2003, pp. 1–6.
9. Z. Lian, A. Godil, and X. Sun, "Visual Similarity Based on 3D Shape Retrieval Using Bag-of-Features," *Shape Modeling Int'l Conf.* (SMI), 2010, pp. 25–36.
10. G.E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*, vol. 18, no. 7, 2006, pp. 1527–1554.

Shuhui Bu is an associate professor of in the School of Aeronautics at Northwestern Polytechnical University, China. His research interests include 3D shape analysis, image processing, pattern recognition, 3D reconstruction, and robotics. Bu has a PhD in computer science from the College of Systems and Information Engineering at the University of Tsukuba, Japan. Contact him at bushuhui@nwpu.edu.cn.

Shaoguang Cheng is a graduate student in the School of Aeronautics at Northwestern Polytechnical University, China. His research interests include artificial intelligence, 3D shape analysis, image processing, and computer vision. Cheng has a BS in electrical engineering and automation from Northwestern Polytechnical University. Contact him at chengshaoguang@mail.nwpu.edu.cn.

Zhenbao Liu is an associate professor in the School of Aeronautics at Northwestern Polytechnical University, China. His research interests include 3D shape analysis, matching, retrieval, and segmentation. Liu has a PhD in computer science from the College of Systems and Information Engineering at the University of Tsukuba, Japan. Contact him at liuzhenbao@nwpu.edu.cn (corresponding author).

Junwei Han is a professor in the School of Automation at Northwestern Polytechnical University, China. His research interests include computer vision and multimedia processing. Han has a PhD in measurement and control technology from Northwestern Polytechnical University. Contact him at jhan@nwpu.edu.cn.