3D Model Annotation based on Semi-Supervised Learning

Kai Zhou†, and Feng Tian†

† School of Computer and Information Technology, Northeast Petroleum University, Daqing, China

Summary
The purpose of annotation for 3D model is that it can automatically list the best suitable labels to describe the 3D models; it is an important part of the text-based 3D model retrieval. The existence of the semantic gap makes the result based on the similarity matching techniques needs to be improved. In order to improve the 3D model annotation performance using a large number of unlabeled samples, we propose a semi-supervised measure learning method to realize the 3D models multiple semantic annotation. A graph-based semi-supervised learning is firstly used to expand the training set, and the semantic words confidence of the models in the extension set is proposed. An improved relevant component analysis method is proposed in this paper to learn a distance measure based on the extended training set. Our approach is introduced to complete multiple semantic annotation task based on the learned distance measure. The test result on the PSB data set have shown that the method making use of the unlabeled samples has achieved a better annotation result when a small amount of labels were given.

Key words:
model automatic annotation, 3D model retrieval, semantic retrieval, metric learning,semi-supervised learning.

1. Introduction

In recent years, 3D scanning equipment, modeling tools and Internet technology have led to a large number of 3D models with widespread, 3D model retrieval become a research hotspot. Several 3D model search engines have been developed. Such as the 3D model search engine at Princeton University [1], the 3D model retrieval system at the National Taiwan University [2], the Ephesus search engine at the National Research Council of Canada [3]. These search engines are all include two search types. One is using traditional text-based retrieval which keywords are extracted from captions, titles, etc. The other type is using content-based retrieval method which search sample is 2d image or 3d object. In contrast, content-based 3D shape retrieval methods, that use shape properties of the 3D models to search for similar models, work better than traditional text-based methods [4]. But compare the 2d image or 3d object, the texture keyword is easier to define and get. The text-based retrieval provides users with a simple and natural interface, so it is friendlier for the user, but the text labels is required. In order to improve the retrieval effectiveness and capture the user's semantic knowledge, the semantic automatic annotation technique has been introduced to the 3D model retrieval broadly in recent years [5-6]. Most current automatic annotation methods need a large number of models hand tagged with text labels, so the training sample size and quality are in high demand [7]. At the same time manually annotation brought tedious workload, which made the label results imperfect, inaccurate and subjective. Figure 1 show some hand tagged models and their labels.

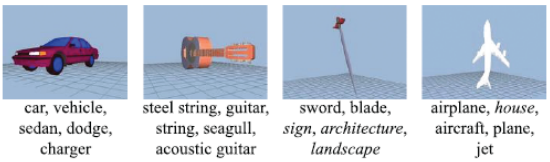

Fig.1. Four hand tagged models.

In this paper, we present a method called 3D model multiple semantic automatic annotation based on semi-supervised metric learning （MS3ML） to label 3D models, which has achieved a better annotation result when a small amount of labels were given. The process flow by MS3ML is shown in Figure 2. The corpus is comprised of a small amount of hand tagged models users provided and a large number of unlabeled models. Firstly, the feature of 3D models was extracted, and the process of dimension deduction is needed. Secondly, we make full use of the unlabeled models to expand the training dataset (known as label propagation) and the label confidence was computed. Thirdly, a new distance metric considered label confidence as well as the correlation between features is learned. Lastly, for each model in unlabeled models collection, we label it by multiple semantic annotation strategy.

2. Graph-based Semantic Label Propagation

Since the amount of labelled models is not sufficient for automatic annotation, we need take full use of labelled and unlabeled models to expand the amount of labelled models. The graph-based semi-supervised learning has become the mainstream of semi-supervised learning because of its efficiency [8-9]. To do this, we use a corpus of known hand tagged models $L=\{(x_1,y_1)\cdots(x_{|L|},y_{|L|})\}$ where x_i denotes the model and y_i denotes i-th model's semantic label collection, $y_i \subset \mathrm{T}$, $T=\{\lambda_1\cdots\lambda_{|T|}\}$ denotes the collection

Manuscript received August 5, 2014
Manuscript revised August 20, 2014

consisting of all labels. $U=\{x_{|L|+1},\cdots,x_n\}$ denotes the unlabeled model. The model x_i is represented by the point x_i in the d dimension feature space.

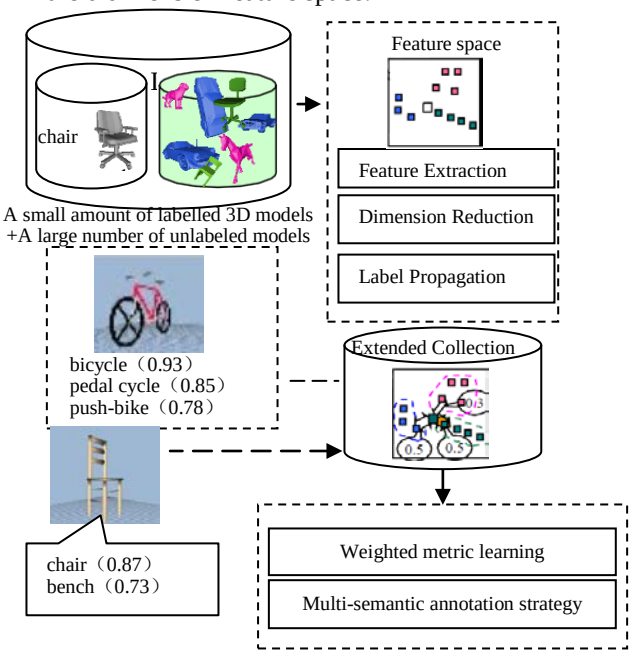

Fig.2. 3D model semantic annotation process of MS3ML

Define a graph, each of its vertices corresponding to each model from $L\cup U$, its weighted edge reflects the similarity between adjacent models. So $n\times n$ similarity matrix W can denotes the graph $G=\{V,E\}$. Each element of the matrix can be formally defined by RBF kernel function as follows:

$$w_{ij}=\exp\left(-\frac{\|x_i-x_j\|^2}{\alpha^2}\right) \tag{1}$$

where w_{ij} denotes the similarity between model x_i and x_j, and $w_{ij}\in(0,1)$. α represents a particular constant. As the label information can be propagated through nodes which are connected by edges of the graph, so a $n\times n$ matrix P can be defined, which represents the edge propagation probability of label information to the neighbour node:

$$P_{ij}=P(j\to i)=\frac{w_{ij}}{\sum_{k=1}^{n}w_{ik}} \tag{2}$$

where P_{ij} denotes probability which x_i learns label information from x_j.

The semantic label of model x_i is expressed by $1\times|T|$ row vector f_i, if $x_i\in L$, the j-th element is defined as follows:

$$f_{ij}=\begin{cases}1, & j\in y_i\\ 0 & j\notin y_i\end{cases} \tag{3}$$

That is, the j-th elements of f_i is 1 if the j-th label in T is one of the models label, and the rest are zero. If $x_i\in U$, $f_{ij}\in[0,1]$. Define the $|L|\times|T|$ matrix f_L which denotes the label semantic matrix, the $|U|\times|T|$ matrix f_U which denotes unlabeled semantic matrix. Define f_X denotes matrix of all the data as follows:

$$f_X=\begin{pmatrix}f_L\\ f_U\end{pmatrix} \tag{4}$$

The data's label is propagated from the neighbors, that is

$$f_X^{(i)}=P\times f_X^{(i-1)} \tag{5}$$

We summarize the standard process of label propagation algorithm as follows:

1. $i=0$, Initialize $f_U^{(i)}=0$;
2. Calculate P;
3. $i=i+1$, get $f_U^{(i)}$ by $f_X^{(i)}=P\times f_X^{(i-1)}$;
4. Repeat step 3 until convergence;
5. Define f_{U_i} as i-th row vector of f_U, each elements of f_{U_i} has been assigned a real-value which is used to measure the confidence of i-th model label in U. We can take columns labels corresponding to the first k largest elements in f_{U_i} as model's semantic labels.

The algorithm steps show that the unlabeled data's label is constantly being updated by the label propagation algorithm, and the labeled data is a starting point, the information of label firstly transferred to the nearest neighbors, then to the secondary neighbors. The final state of label propagation is all the vectors of the unlabeled data are no longer changed, that is semantic labels achieves a smooth distribution in all the unlabeled data. Thus each model has the first k semantic labels. We expands the manually labeled data set L to $L\cup U$, meanwhile for each label we assigned a confidence value which we interpret as the probability that the label is a relevant to the model. So the relevance of each model in $L\cup U$ to each label can be described by the triple like this (xi, 'airplane', 0.83).

3. Weighted Metric Learning

The above method can extend the labelled data set, and we get its label confidence approximately. Now we can learn a new distance metric considered the size of label information as well as the correlation between features with the data in the extended labelled dataset $L\cup U$. RCA

(Relevant Component Analysis) is a simple and effective distance metric learning method. It can learn a global linear transformation in the same class constraints that users provided. In pattern recognition field its performance is better than the usual Euclidean distance and other confidence we got are a guarantee of the algorithm's validity.

We firstly normalize the label confidence of each label in T, and then get a $|T|$-dimensional diagonal matrix of confidence:

$$W = Diagonal[w_1, w_2, w_3, ..., w_{|T|}]$$

where w_i is mean confidence of all the models described by i-th label in T.

So we can use weighted covariance matrix instead of centralized covariance matrix of RCA. We summarize the process of the MS3ML algorithm as follows:

(i) Given the data set $L = \{(x_1, y_1)...(x_{|L|}, y_{|L|})\}$, which is comprised of $|L|$ models, we propagate the label of the model to the unlabeled model set U, so we get the label's confidence of each model which has been propagated (see section 3).

(ii) Each label's confidence are normalized, and a $|T| \times |T|$ diagonal matrix of confidence is generated, the weighted covariance matrix of all the labeled model are calculated:

$$C = \frac{1}{n}\sum_{c=1}^{|T|}\sum_{i=1}^{n_i}(x_{c,i} - x_{c-mean})W(x_{c,i} - x_{c-mean})^T \quad (7)$$

where $x_{c,i}$ denotes *i-th* model in feature space described by *c-th* label. x_{c-mean} is mean point in feature space described by *c-th* label.

(iii) Calculate C^{-1} as a mahalanobis distance metric:

$$d_{weighted-RCA}(x_1, x_2) = (x_1 - x_2)^T C^{-1}(x_1 - x_2) \quad (8)$$

(iv) For each model in unlabeled models collection, we label it by multiple semantic annotation strategy. (see section 5).

4. Multiple Semantic Annotation For The Unlabeled Models

Given an unlabeled 3D model X_{new}, we wish assign labels from the set of all possible labels $T = \{\lambda_1...\lambda_{|T|}\}$ to X_{new}. Specifically, for each label we wish to assign a confidence value which we interpret as the probability that λ_i is a relevant label for X_{new}. So we start with a geometric shape similarity metric and find the neighbors of X_{new} within some distance threshold. Note that the distance threshold is allowed to be a function of the model, which allows for adaptively defining the threshold based on the density of models in a given portion of the descriptor space. We take distance metrics [10]. But we found that when the amount of labelled information is insufficient, the result got from traditional RCA will bias. So we propose a method called weighted RCA. The extended labelled dataset and labelling

$$P(X_{new} \approx X_{neighbour-i}) = (1 - d_{weighted-RCA}(X_{new}, X_{neighbour-i}))^2 \quad (9)$$

to be an estimate of the probability that X_{new} and $X_{neighbour-i}$ represent the same type of model and therefore should have similar text labels. Then given our unlabelled model X_{new}, a possible text label λ_i, and a neighbor $X_{neighbour-i}$ from the extended labeled data set $L \cup U$ (see section 3), the probability that X_{new} should have the label is

$$C(\lambda^i, X_{new}) = P(X_{new} \approx X_{neighbour-i}) \wedge C(\lambda^i, X_{neighbour-i}) \quad (10)$$

Where $C(\lambda^i, X_{new})$ denotes the confidence of label λ_i intuitively this means that the probability that λ_i is appropriate for X_{new} is the probability that it is appropriate for $X_{neighbour-i}$ and that X_{new} and $X_{neighbour-i}$ are similar enough to share labels. $C(\lambda^i, X_{new})$ can be thought of as measuring how much we trust the original annotation on $X_{neighbour-i}$. When considered over the full set of k neighbors this generalizes to

$$C(\lambda^i, X_{new}) = \bigcup_{j=1}^{k} P(X_{new} \approx X_{neighbour-j}) \wedge C(\lambda^i, X_{neighbour-j}) \quad (11)$$

By analogy to the *TF-IDF* method from text processing we reweight these probabilities such that:

$$C_{tf-idf}(\lambda^i, X_{new}) = \frac{C(\lambda^i, X_{new})}{\sum_j C(\lambda^j, X_{new})} \cdot \log \frac{|L \cup U|}{\sum_{X_k \in L \cup U} C(\lambda^i, X_k)} \quad (12)$$

For each unlabeled model, we get a vector of probabilities for each semantic labels. We choose the *TOP-N* labels to describe the model.

5. Experiment

To evaluate the proposed method, our experiments were performed on a database containing 1125 3D models, which were collected from the Princeton Shape Benchmark (PSB). In order to evaluate the methods described in this paper, 725 were semantically hand tagged with text labels. In the experiment, this paper mainly uses

the depth buffer method to extract the 3D models' feature (438-dimensional feature vector) in the model base [11]. We performed a PCA over the descriptors, and kept only the top 20 dimensions.

In this paper, we use "Average Precision" VS "Percentage of each tag labeled" to evaluate both automatic labeling process and retrieval process, figure 3 lists the average retrieval precision of five times. These types of labeling methods, including: Euclidean distance metric method, a typical supervised classification learning method (SVM algorithm and the Euclidean distance), RCA distance metric method (RCA algorithm and mahalanobis distance) and MS3ML.

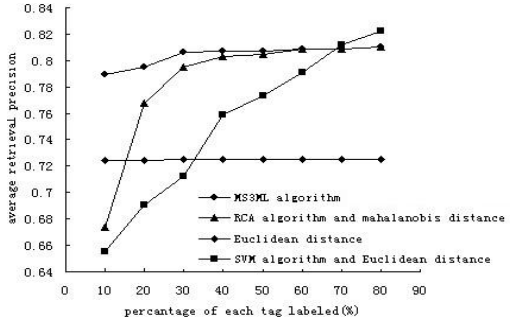

Fig.3. Comparison of the average retrieval precision

Figure 3 shows that the proposed method has the higher labelling precision when there is a small amount of label information. Among them, the kernel function of the supervised labelling method SVM adopted RBF kernel [12]; the distance metric function used Euclidean distance. Since SVM requires a large number of training data, so if we select a few data sets for training, the labelling result was not be accurate and led to low retrieval precision. In order to test the validity of the proposed method on a small labelled information. Table 1 respectively shows the average retrieval efficiency of various methods in the case of very few labelled data (label 1, 2, 3 and 4 models for each label), and only the first 16 retrieval results will be taken into account.

Table1. Comparison of the retrieval effectiveness of several 3D model retrieval methods with a small amount of labels

methods		**Labeled models per label**	**precision (%)**	**recall (%)**
Supervised method	SVM and Euclidean distance	1	17.31	5.51
		2	34.47	10.64
		3	49.52	16.82
		4	64.67	22.93
Semi-supervised method	RCA	1	38.92	13.12
		2	44.02	14.65
		3	46.41	15.92
		4	66.51	23.42
	MS3ML	1	74.11	26.51
		2	75.53	26.97
		3	77.55	27.78
		4	78.07	28.08

Table 1 shows this proposed semi-supervised distance metric learning methods has a better retrieval results in the case of very few labelled data information.

6. Conclusion And Future Work

In this paper, we have proposed a novel method for multiple semantic automatic labeling of 3D models by semi-supervised metric learning (MS3ML). The method acquires a small amount of hand tagged information provided by users, and the semi-supervised semantic label propagation takes full use of unlabeled models to expand the training dataset. The expanded collection increase in the number of labeled models; meanwhile labeling confidence we got can describe the semantic relevance of the label, on the basis of the above two points, Weighted-RCA method can effectively resolve the traditional RCA learning bias caused by the insufficient amount of labeled data or inaccurate labeling information. The result on the Princeton Shape Benchmark shows that MS3ML get a better retrieval results and performance, so the method not only reduce hand tagged information, but also improves the retrieval accuracy in the case of very few labelled data information.

In addition, in our experiments, we observed that this method also has some limitations, because the algorithm requires that the small amount of labels provided is correct, if there are errors in these labels, it will reduce the efficiency of the algorithm. In future work we will focus on how to improve the robustness of the algorithm and perfect the annotation result by relevance feedback from search result.

Acknowledgments

This work is supported by Youth Foundation of Northeast Petroleum University (NO: 2013NQ120, 2012QN117), Scientific Research Fund of Heilongjiang Provincial Education Department (NO: 12521055). We also would like to thank the anonymous reviewers for their helpful comments and suggestions.

References

[1] S.Philip, M.Patrick, K.Michael. "The Princeton Shape Benchmark," Shape Modeling International, 2004,pp.388-399

[2] Chen Ding-yun, Tian Xiao-per, Shen Yu-te. "On visual similarity based 3D model retrieval". Eurographics, 2003, pp. 22-26.

[3] Paquet E, Murching A, Naveen T, Tabatabai A, Rioux M. "Description of shape information for 2-D and 3-D objects". Signal Process Image Commun. 2000. 16:pp.103–122

[4] Min P, Kazhdan M, Funkhouser T. A comparison of text and shape matching for retrieval of online 3D models. In: Proc. European conference on digital libraries, 2004,pp.209–220
[5] Meng Z, Atta B. “Semantic-associative visual content labeling and retrieval: A multimodal approach”.Signal Processing: Image Communication. 2007,pp.569-582
[6] T. Funkhouser, P. Min, M. Kazhdan, J. Chen, A. Halderman, D. Dobkin, and D. Jacobs. “A Search Engine for 3D Models”. ACM Transactions on Graphics, 22(1), 2003,pp.83-105.
[7] Ritendra Datta , Weina Ge , Jia Li , James Z. Wang, “Toward bridging the annotation-retrieval gap in image search by a generative modeling approach,” Proceedings of the 14th annual ACM international conference on Multimedia, 2006,pp.23-27, Santa Barbara, CA, USA [doi:10.1145/1180639.1180856]
[8] D.Zhou, O.Bousquet, T.Lal, J. Weston, B.Scholkopf. “Ranking on data manifolds,” Proceedings of the 18th Annual Conference on Neural Information Processing System. 2003,pp.169-176
[9] X.Zhu, Z.Ghahramani,and J.Lafferty. “Semi-supervised learning using Gaussian fields and harmonic functions,” Proceedings of International Conference on Machine Learnings,2003
[10] Noam S, Tomer H, Daphna W. “Adjustment learning and relevant component analysis,” Proc. European Conference of Computer Vision (ECCV), Copenhagen, 2002,pp.776-790.
[11] M.Heczko, D.Keim, D.Saupe. “Methods for similarity search on 3D databases,” Datenbank-Spektrum (In German) , 2002, 2(2),pp.54-63
[12] HOU S, LOU K, RAMANIK. “SVM based semantic clustering and retrieva l of a 3D m ode l database.” Journal of Computer Aided Design and Application, 2005,pp. 155-164

Kai Zhou received the M.E. degrees from Northeast Petroleum University in 2006. She is research assistant since 2006 and is working as a lecturer from 2009 in the Department of Computer and Information Technology, Northeast Petroleum University. Her research interest includes pattern recognition, virtual reality.

Feng Tian born in 1980. PhD and associate professor. Received his PhD degree in computer application technologies from the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University in 2014. His main research interests include image annotation, image tagging, cross media analysis, multimedia mining, 3D model retrieval, virtual reality and pattern recognition.