3D SHAPE SIMILARITY USING VECTORS OF LOCALLY AGGREGATED TENSORS

Hedi Tabia[⋆]   *David Picard*[⋆]   *Hamid Laga*[‡∗]   *Philippe-Henri Gosselin*[⋆†]

[⋆] ETIS/ENSEA, University of Cergy-Pontoise, CNRS, UMR 8051, France
[‡]Phenomics and Bioinformatics Research Centre, University of South Australia, Australia
[∗]Australian Centre of Plant Functional Genomics (ACPFG)
[†]INRIA Rennes Bretagne Atlantique, France

ABSTRACT

In this paper, we present an efficient 3D object retrieval method invariant to scale, orientation and pose. Our approach is based on the dense extraction of discriminative local descriptors extracted from 2D views. We aggregate the descriptors into a single vector signature using tensor products. The similarity between 3D models can then be efficiently computed with a simple dot product. Experiments on the SHREC12 commonly-used benchmark demonstrate that our approach obtains superior performance in searching for generic shapes.

Index Terms— 3D Shape retrieval, Depth images, Bag of Features.

1. INTRODUCTION

Large databases of 3D models available in the public domain have created the need for shape analysis and retrieval algorithms able to find similar shapes in the same way a search engine responds to text and image queries. State-of-the-art 3D retrieval methods [1] often follow three main steps: feature extraction and description, descriptors aggregation into compact signatures, and learning of the similarity metric.

Feature-based methods, that have been widely adopted by the computer vision community for image retrieval and scene understanding, are gaining popularity in 3D shape analysis and retrieval [2, 3, 4, 5, 6, 7]. One of the strengths of feature-based approaches in 3D shape analysis is that they represent an object as a collection of primitive elements, called visual features, and use the Bag of Features (BoF) paradigm for aggregation. This paradigm makes use of feature detectors and descriptors to build a visual vocabulary. Each object is then represented with a vector of frequencies of occurrences of visual features. These vector representations can be compared with standard distances, and subsequently be used by robust classification methods such as Support Vector Machines.

The challenge in using dictionary based approaches is to find a good tradeoff between discrimination power, the size of the dictionary and therefore the size of the descriptor, and the scalability to large databases. Standard BoF techniques [8, 9], initially proposed for image indexing and recently used in 3D retrieval [10] require large code books, usually in the scale of 10k to achieve acceptable performance on relatively small databases of 3D models (1200 models). Several schemes have been proposed to improve the performance. This includes the coding/pooling schemes using the Locality-constrained Linear Coding (LCC) [11], which uses the same code size but achieves better performance in image categorization. Other methods, such as the Fisher Vector (FV) [12], use model deviation. This technique models descriptors as a Gaussian Mixture, and computes an image index as the deviation between the database model and the image model. The FV achieves better results than standard BoF approaches [9], but at a cost of larger index size and search time. Jegou et al. [13] proposed approximation and compression techniques in order to use FV-based BoF on large image datasets. Another strategy for building efficient signatures is to linearize the kernel function on bags [14]. A Kernel function on bags aims at computing a similarity analog to vote-based systems, but with respect to Mercer's condition. Their huge computational complexity can be highly reduced using linearization techniques based on tensors, such as Vector of Locally Aggregated Tensors (VLAT) [15] whose performances are comparable to FV.

In this paper we bring the spirit of feature-based computer vision approaches to the problem of 3D shape indexation and retrieval. Our approach starts with the extraction of dense descriptors from each shape. The advantage of using dense descriptors, in contrast to few descriptors computed at sparse locations, is well acknowledged by the computer vision community as it enables efficient image classification. The gain in performance comes at the expense of significant increase in computation time and memory requirement. Our main contribution is a new procedure for descriptor aggregation into compact signatures using the VLAT method. It sums tensor products of the local descriptors that showed to be a good approximation of insightful similarity measures between descriptors. Our experiments on standard 3D retrieval benchmarks show that our approach achieves 83% in retrieval performance, outperforming standard BoF approaches with more than 49%. We details the approach in Section 2 and validate it with experiments on standard benchmarks in Section 3.

2. METHOD OVERVIEW

Figure 1 gives an overview of the proposed method. First in a pre-processing step, we normalize the models to ensure that the extracted descriptors are invariant to translation and scale. Then, we render depth maps of the object from n views uniformly sampled on the surface of a bounding unit sphere. We represent each depth map as a collection of dense Histogram of Oriented Gradients (HoG) descriptors. HoG have the advantage of being compact and easy to compute. We then build the shape signature by aggregating the descriptors using the BoF paradigm. Standard BoF approaches would result in large dictionaries. In this paper, we explore for the first time the usage of Vector of Locally Aggregated Tensors (VLAT) for 3D model indexation. VLAT have been originally proposed for image analysis [14] and proven to be compact and efficient for large-scale image retrieval. In this paper we further reduce their size using Principal Component Analysis (PCA) on the VLAT vectors, while improving further their discrimination power. Our method has several advantages. It is invariant to rigid transformations and some articu-
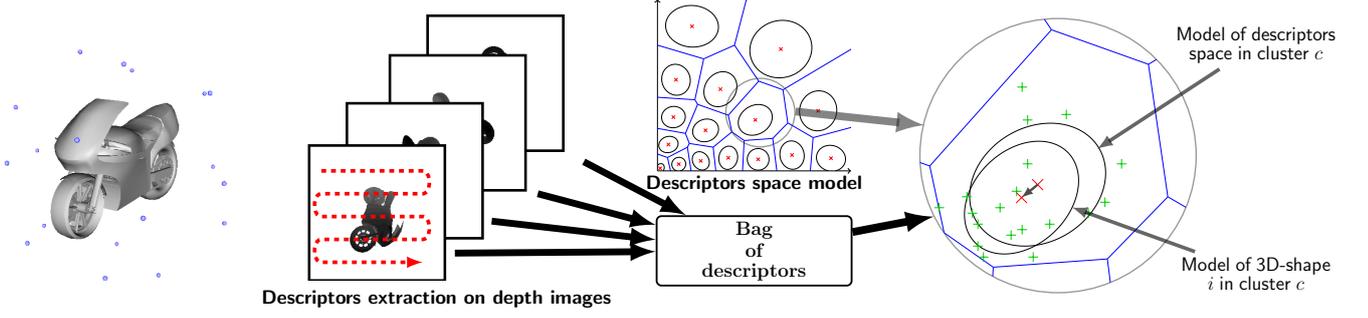
Fig. 1. Method overview: First we compute depth images of the model captured from cameras localized on the unit sphere. Then, local features (*e.g.* HoG) are extracted from the images. Finally, we compute the VLAT signature using deviations between covariance matrices of the codebook and covariance matrices of the descriptors.

lated deformations and robust to geometrical and topological noise. It is robust to the level of tessellation of 3D models. It can handle any type of 3D shape representations as long as depth images can be rendered. The important deviation from the state-of-the-art is that it relies on a compact dictionary (around 512 codewords) but achieves significantly better performance than existing methods as shown in Tables 1 and 2.

2.1. Model normalization

Similar to many 3D retrieval algorithms, our approach is invariant to rigid transformations such translations, scalings, and rotations. In a pre-processing step, we normalize every 3D object for translation and scale by translating it to its center of mass and scaling it such that the radius of the smallest enclosing sphere is one [3]. The use of the smallest enclosing sphere has several advantages: it is fast to compute, it allows the maximization of the model size inside the unit sphere.

Here, we do not perform the pose normalization for rotation. Our approach is based on dense local descriptors and BoF paradigms and hence it is invariant to rotations.

2.2. Dense feature extraction

We represent a 3D object with a set of depth maps captured by virtual cameras distributed uniformly around the object, similar to [16] (see also Figure 2). In order to capture all the important features of the object, we use a large number of views (80 in our implementation) and capture depth images of size 256×256. From the depth images, we extract a dense set of HoG descriptors on a dense regular grid. Every two pixels we compute one HoG descriptor at four different scales (16×16, 24×24, 32×32 and 40×40 pixels). We then obtain a large unordered set of local descriptors. Let $\mathbf{B}_i = \{\mathbf{b}_{ir}, r = 1 \ldots n\}$ be the set of descriptors extracted from the depth images of model i. Unlike [3], we do not maintain a separate set of descriptors for each view. Instead, we put all the descriptors in a single bag and use them to build our compact dictionary. This makes the descriptor invariant to rotations of the 3D model. Also, given a sufficiently large number of views and a densely sampled HoG descriptors, we insure that the most important features of an object are captured by the representation.

One method to map the set of descriptors into a single vector is the Bag of Features (BoF) [8]. It involves the construction of a visual codebook (visual words) and the count of occurrences of these words in the 3D model. When using large and dense set of visual descriptors, large dictionaries from 4k to 64k visual words are required to get good performances with BoF. We propose to use Vectors of Locally Aggregated Tensors (VLAT) recently proposed in [14] for image analysis to build very discriminative shape signatures using a small dictionary size.

2.3. VLAT: Tensor based aggregation

Let $\mathbf{B}_i = \{\mathbf{b}_{ir}\}$ and $\mathbf{B}_j = \{\mathbf{b}_{jr}\}, r = 1 \ldots n$, be two bags of features representing the sets of descriptors in two 3D shapes i and j. An effective method to compute the similarity between two bags is based on kernel functions. Thanks to mathematical properties like Mercer conditions, these kernel functions on bags can be used with many powerful kernel-based learning techniques. The novelty in this paper is that we consider a kernel function on bags for each cluster c:

$$K(\mathbf{B}_i, \mathbf{B}_j) = \sum_c K_B(\mathbf{B}_{ic}, \mathbf{B}_{jc}) \qquad (1)$$

with $\mathbf{B}_{ic} = \{\mathbf{b}_{rci}\}_r$ the descriptors of model i that belong to cluster c. K_B is a kernel on the bags $\mathbf{B}_{ic}$ and $\mathbf{B}_{jc}$, which we define as the sum of Gaussian kernels on each pair of descriptors belonging to $\mathbf{B}_{ic}$ and $\mathbf{B}_{jc}$.

$$K_B(\mathbf{B}_{ic}, \mathbf{B}_{jc}) = \sum_{r,s} e^{-\frac{1}{\sigma^2} ||\mathbf{b}_{rci} - \mathbf{b}_{scj}||^2} \qquad (2)$$

Computing a Gaussian kernel on each pair of descriptors is computationally prohibitive, especially for large bags ($O(n^2)$ for a dictionary of size n). Thus our second idea is to linearize the Kernel functions on bags. To do so, we first normalize all the descriptors to have a unit length. Then we expand it using Taylor series, and finally linearize the kernel function using tensor products:

$$
\begin{aligned}
K_B(\mathbf{B}_{ic}, \mathbf{B}_{jc}) &= e^{-\frac{2}{\sigma^2}} \sum_{r,s} e^{\frac{2\langle \mathbf{b}_{rci}, \mathbf{b}_{scj}\rangle}{\sigma^2}} \\
&= e^{-\frac{2}{\sigma^2}} \sum_{r,s} \sum_p \frac{\alpha_p}{\sigma^{2p}} \langle \mathbf{b}_{rci}, \mathbf{b}_{scj}\rangle^p \\
&= e^{-\frac{2}{\sigma^2}} \sum_p \frac{\alpha_p}{\sigma^{2p}} \langle \sum_r \otimes^p \mathbf{b}_{rci}, \sum_s \otimes^p \mathbf{b}_{scj}\rangle
\end{aligned}
$$

with $\otimes^p \mathbf{x}$ the tensor product of order p of the vector $\mathbf{x}$.
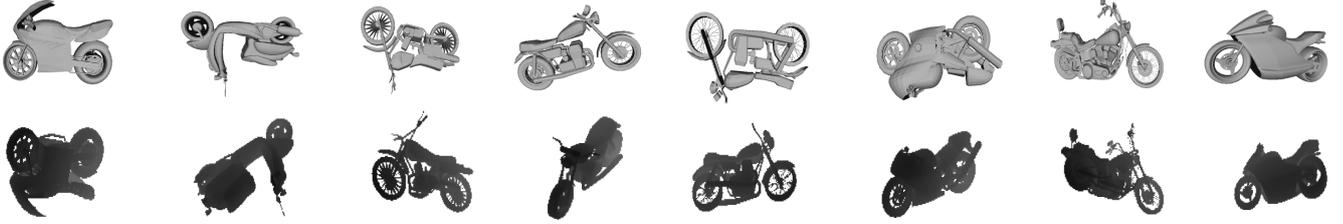
Fig. 2. Example of 3D models from the motorcycle category from the SHREC12 data-set. The top row shows renderings of the models, while the bottom row shows depth images extracted randomly for the corresponding models.

When stopped at $p = 2$, the expansion corresponds to the second order statistics of the set $\mathbf{B}_{ic}$. To extends the range of the similarity measure, we propose to center the obtained tensors by the mean tensors of each cluster c. The mean and tensors of the cluster have been learned during the dictionary construction. We proceed as follows;

- First, we compute the mean descriptor μ_c and mean tensor matrix $\mathcal{T}_c$ of each cluster c:

$$\mu_c = \frac{1}{|c|} \sum_i \sum_r \mathbf{b}_{ric} \qquad (3)$$

$$\mathcal{T}_c = \frac{1}{|c|} \sum_i \sum_r (\mathbf{b}_{ric} - \mu_c)(\mathbf{b}_{ric} - \mu_c)^T \qquad (4)$$

where $|c|$ is the number of descriptors in cluster c and $\mathbf{b}_{rci}$ are the descriptors of model i that belong to c. The quantities μ_c and $\mathcal{T}_c$ provide statistical summaries of the shape cluster and capture the main variabilities.

- Next, for every 3D model i we compute one feature per cluster c. We compute a tensor $\mathcal{T}_{ic}$ as an aggregation of the centered tensors of centered descriptors:

$$\mathcal{T}_{ic} = \sum_r (\mathbf{b}_{rci} - \boldsymbol{\mu}_c)(\mathbf{b}_{rci} - \boldsymbol{\mu}_c)^\top - \mathcal{T}_c \qquad (5)$$

where $\boldsymbol{\mu}_c$ is the center of cluster c. We flatten the matrix $\mathcal{T}_{ic}$ into a feature vector $\mathbf{x}_{ic}$, the descriptor of the 3D shape i with respect to cluster c.

- Finally, we concatenate all the vectors $\mathbf{x}_{ic}$ of object i to form a single feature $\mathbf{X}_i = \{\mathbf{x}_{ic}\}, c = 1 \ldots n$, where n is the size of the dictionary. $\mathbf{X}_i$ is called the VLAT signature of model i.

Note that the size of VLAT features is $n \times d \times d$, where n is the number of clusters and d is the dimension of the descriptor. Its dimensionality can be significantly reduced using Principal Component Analysis (PCA) on some training data.

2.4. PCA-based VLAT (PVLAT)

We propose an efficient method for reducing the dimensionality of the shape descriptor while preserving its high discrimination power. For a given cluster of shapes c, we perform a Principal Component Analysis on the descriptors that belong to this cluster. We compute the leading eigenvalues λ_l and eigenvectors Λ_l of the tensor matrix $\mathcal{T}_c$. Then we project the centered descriptors belonging to c on the eigenvectors

$$\mathbf{b}'_{rci} = \Lambda^\top (\mathbf{b}_{rci} - \boldsymbol{\mu}_c). \qquad (6)$$

	D	NN	1-Tier	2-Tier	e-Measure	DCG
BOF	256	0.3267	0.1591	0.2297	0.1560	0.4722
BOF	512	0.3392	0.1643	0.2431	0.1647	0.4814
BOF	4096	0.5908	0.2873	0.3889	0.2706	0.6109
BOF	16384	0.6858	0.3436	0.4637	0.3240	0.6688
VLAT	**256**	**0.7733**	**0.4088**	**0.5021**	**0.3467**	**0.7017**
VLAT	**512**	**0.7767**	**0.4074**	**0.5158**	**0.3654**	**0.7173**
PVLAT	**256**	**0.8125**	**0.4647**	**0.5953**	**0.4207**	**0.7661**
PVLAT	**512**	**0.8342**	**0.4844**	**0.6234**	**0.4421**	**0.7807**

Table 1. Comparison with Bag of Features. The column D refers to the dictionary size.

where Λ is a matrix whose columns are the leading eigenvectors Λ_l. The new VLAT signature, called PVLAT, are computed on reduced descriptors $\mathbf{b}'_{rci}$. This optimization has the very interesting property that most of the variance is concentrated among the first dimensions of each vector.

3. EXPERIMENTS AND RESULTS

To assess the performance of our method we use the SHREC 2012 Generic 3D Shape Retrieval benchmark [10]. The dataset contains 1200 models, classified into 60 categories mainly based on visual similarity. In the implementation of our method, we generated 80 depth maps. For each depth map, we extract HoG descriptors on a dense grid, one HoG feature every two pixels, with four different scales.

We run three different experiments. In the first one (Table 1), we analyze the effect of the dictionary size on the performance of the proposed compact signature. In the second experiment (Table 1), we compare the performance of our method to standard bag of features techniques. Finally, we compare our methods to state-of-the-art methods benchmarked in SHREC 2012 [10]. We use five metrics to evaluate the proposed approach: the Nearest Neighbor (NN), First-tier (1-Tier), Second-tier (2-Tier), E-Measures and Discounted Cumulative Gain (DCG). We also report the precision-recall graph.

Table 1 shows the performance of our method when using a dictionary of size 256 and 512. As we can see, both proposed signatures are fairly stable with respect to the codebook size. Compared to standard BoF approach (Table 1), we can see clearly that our method performs significantly better, both in retrieval performance and memory efficiency. With only a dictionary of size 512 our PVLAT method achieved 0.83 in the Nearest Neighbor measure against 0.68 for a standard BoF with a dictionary of size 16384.

Table 2 shows a comparative evaluation of our method and five other methods presented in [10]. As one can see our method

Methods	NN	1-Tier	2-Tier	e-Measure	DCG
DG1SIFT [10]	0.879	0.661	0.799	0.576	0.871
PVLAT	**0.834**	**0.484**	**0.623**	**0.442**	**0.780**
DVD+DB [10]	0.831	0.496	0.634	0.450	0.785
ZFDR [10]	0.818	0.491	0.621	0.442	0.776
VLAT	**0.776**	**0.407**	**0.515**	**0.365**	**0.717**
3DSP_L3_200_hik [10]	0.708	0.361	0.481	0.335	0.679
LSD-sum [10]	0.517	0.232	0.327	0.224	0.565

Table 2. Comparison with state-of-the-art methods as reported in the shape retrieval context SHREC 2012 [10] .
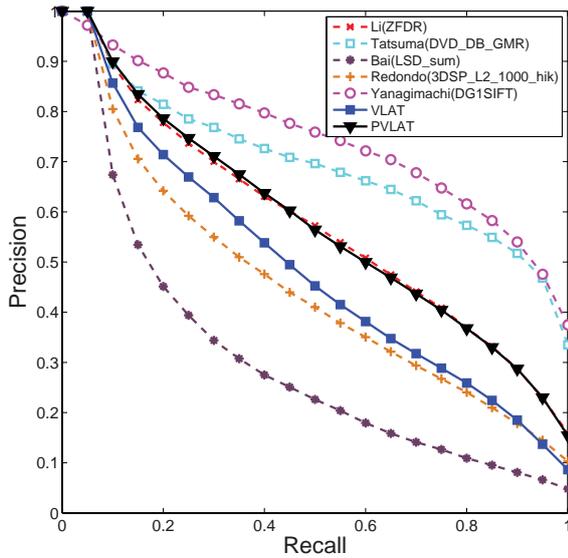

Fig. 3. Precision recall graph for our approach on the SHREC 2012 dataset compared to state-of-the-art graphs.

performs better than LSD-sum, 3DSP, ZFDR and DVD+DB. The DG1SIFT performs the best. Note that DG1SIFT combines three different methods for descriptor sampling: dense regular sampling, random sampling, and one global SIFT descriptor. DG1SIFT uses also the BoF paradigm, with a vocabulary size exceeding 13k, which is far higher than the approach we proposed in this paper. Furthermore, our method uses a linear metric (dot product), unlike the methods reported in [10], which rely on computationally intensive non-linear similarity measure. These experiments show that our approach achieves good performances on a standard 3D database with respect to its simplicity.

Figure 3 presents the Precision vs Recall plots of our method and the state of the art methods from [10]. It is interesting to notice that our method and the ZFDR one present quite comparable performances, however our method's precision is slightly higher for low recall values. Our method clearly outperforms LSD-sum and 3DSP methods.

We also compare our proposed PVLAT signature against the standard VLAT. Table 2 and Figure 3 show that the PVLAT achieves significantly better performance while significantly reducing the size of the signature.

4. CONCLUSION

We proposed a new 3D model indexing method using vectors of locally aggregated tensors on local descriptors densely extracted from depth views. Our signature uses a single descriptor and a linear metric. The method showed very high retrieval accuracy on a standard 3D shape retrieval benchmark, exceeding previous state-of-the-art BoF approaches. These two properties make the proposed method suitable for large-scale retrieval of 3D models, on which we plan to test in the future.

5. REFERENCES

[1] Johan W. H. Tangelder and Remco C. Veltkamp, "A survey of content based 3d shape retrieval methods," in *Multimedia Tools and Applications*, 2008, pp. 441–471.

[2] Alexander M. Bronstein, Michael M. Bronstein, Leonidas J. Guibas, and Maks Ovsjanikov, "Shape google: Geometric words and expressions for invariant shape retrieval," *ACM Trans. Graph.*, vol. 30, 2011.

[3] Zhouhui Lian, Afzal Godil, and Xianfang Sun, "Visual similarity based 3d shape retrieval using bag-of-features," in *Proceedings of the 2010 Shape Modeling International Conference*, Washington, DC, USA, 2010.

[4] Hedi Tabia, Mohamed Daoudi, Olivier Colot, and Jean-Philippe Vandeborre, "Three-dimensional object retrieval based on vector quantization of invariant descriptors," *SPIE Journal of Electronic Imaging*, vol. 21, no. 2, pp. 023011–1– 023011–8, April-June 2012.

[5] Hedi Tabia, Mohamed Daoudi, Jean-Philippe Vandeborre, and Olivier Collot, "Local visual patch for 3D shape retrieval," in *ACM International Workshop on 3D Object Retrieval (in conjunction with ACM Multimedia 2010)*, Firenze, Italy, October 25 2010.

[6] Hamid Laga, Masayuki Nakajima, and Kunihiro Chihara, "Discriminative spherical wavelet features for content-based 3d model retrieval," *International Journal of Shape Modeling (IJSM)*, vol. 13, no. 1, pp. 51–72, June 2007.

[7] Hamid Laga, Hiroki Takahashi, and Masayuki Nakajima, "Spherical wavelet descriptors for content-based 3d model retrieval," in *SMI '06: Proceedings of the IEEE International Conference on Shape Modeling and Applications 2006 (SMI'06)*, Washington, DC, USA, 2006, pp. 75–85, IEEE Computer Society.

[8] Josef Sivic and Andrew Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proceedings of ICCV '03*, Washington, DC, USA, 2003, pp. 1470–.

[9] Ken Chatfield, Victor Lempitsky, Andrea Vedaldi, and Andrew Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *BMVC*, 2011, vol. 76, pp. 1–12.

[10] B. Li, Afzal Godil, Masaki Aono, X. Bai, Takahiko Furuya, L. Li, Roberto Javier Lpez-Sastre, Henry Johan, Ryutarou Ohbuchi, Carolina Redondo-Cabrera, Atsushi Tatsuma, Tomohiro Yanagimachi, and S. Zhang, "Shrec'12 track: Generic 3d shape retrieval.," in *3DOR*. 2012, pp. 119–126, Eurographics Association.

[11] Jingjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010, pp. 3360–3367.

[12] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, "Improving the fisher kernel for large-scale image classification," in *ECCV*, 2010, pp. 143–156.

[13] Herve Jegou, Florent Perronnin, Matthijs Douze, Jorge Sanchez, Patrick Perez, and Cordelia Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, 2012.

[14] Romain Negrel, David Picard, and Philippe-Henry Gosselin, "Compact tensor based image representation for similarity search," in *ICIP*, Orlando, Florida, USA, September 2012.

[15] Romain Negrel, David Picard, and Philippe-Henry Gosselin, "Using spatial pyramids with compacted vlat for image categorization," in *ICPR*, Tsukuba Science City, Japan, November 2012.

[16] Hamid Laga and Masayuki Nakajima, "Supervised learning of salient 2D views of 3D models," *The Journal of the Society for Art and Science*, vol. 7, no. 4, pp. 124 –131, December 2008.