Vis Comput (2012) 28:279–287
DOI 10.1007/s00371-011-0638-z

ORIGINAL ARTICLE

Web-image driven best views of 3D shapes

Hong Liu · Lei Zhang · Hua Huang

Published online: 15 September 2011
© Springer-Verlag 2011

Abstract The rapid advance of the Internet provides available huge database of web images. In this paper, we introduce a novel approach for automatically computing the best views of 3D shapes based on their web images. Best view selection is generally an intuitive task of getting the most information of a 3D shape. The novelty of our approach is to directly explore human perception on observing 3D shapes from the relevant web images. Those images are captured from biased views of different people, thus sufficiently reflecting view choice when observing the 3D shapes. By collecting web images possibly captured from the similar views, the best view is selected as the one possessing the most web images. We experiment our method with the shapes in Princeton Shape Benchmark (PSB), as well make comparisons with traditional geometric descriptor based approaches. The results demonstrate that our method is not only robust but also able to produce more canonical views in accordance with human perception.

Keywords Best view · Web images · Silhouette · Curvature · Saliency

1 Introduction

When a 3D shape is observed from different views, it might present dramatically varied appearance and convey different visual impression about the 3D shape. To recognize and understand the 3D shape, it is better to seek for a "canonical" view to see as much information as possible, which is regarded as the problem of best view selection. Various applications involve best view selection as an important step to conduct convincing results, such as image based modeling [1], shape recognition and classification [2], thumbnails in 3D models searching [3], and so on.

Generally, there is no consensus about what is a good view for a 3D shape. Under different conditions, people would always use their own perception or definition to meet with intended applications to select the best view. However, Blanz et al. [4] proposed four attributes that approve the "canonical" views after learning computer-graphics psychophysics: goodness for recognition, familiarity, functionality and aesthetic criteria, and stated the best view appears to be largely familiarity based and clearly affected by geometrical properties. Based on this research, selection of the best view of a 3D shape turns to finding the view that providing us the most intuitive information about the 3D shape [2, 3, 5]. In the practical computation, some descriptors, like curvature, topology, or silhouette entropy, are applied to represent the intuitive information, so the best view problem is obtained by maximizing the visibility of these descriptors. These descriptors may work well based on the mathematical soundness, but as shown in Fig. 9, for some shapes, they may not meet the intuitive observation by human perception.

Since the goodness of best views of 3D shapes is perceptually judged, the straightforward yet credible solution is opinion polls on the best views. However, there are billions of people on the earth, and querying for opinions on the best view choice of a 3D shape is impossible. Fortunately, the Internet becomes an easy-to-access platform where people can publish their photographed images. For example, there exist more than 4 billion web images shared in the famous website *Flickr*, and 10 billion images can be searched by *Google*. More importantly, those web images contain view information about the captured 3D models, as people tend

H. Liu · L. Zhang (✉) · H. Huang
School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China
e-mail: cgzhanglei@gmail.com

to choose their favorite views in photographing. The motivation of our approach is to make use of the web-image database to investigate the preference of view selection for 3D shapes involved in the images.

The main contribution of our paper is a web-image driven approach to best view computation, which is solved directly from human perception on the 3D shapes. To facilitate the web-image driven selection, a novel similarity measure is introduced by combining both 3D-2D shape and saliency features. Based on the similarity measure, each web image tallies a vote to the candidate view with the best matching. Our approach selects the best view with the most votes from web images, which can reflect the human perception on 3D shapes completely and advocated by the user study.

2 Related work

Since best view can be cast as the most informative one, much research work employs information theory to solve the problem. Vázquez et al. [6, 7] initiate viewpoint entropy to compute the best views, and use the projected area of all the visible triangles as entropy to measure the best view with maximum relative projective area. Page et al. [8] turn to curvature to define the entropy, which describes the 3D shape with more geometric details. However, curvature cannot totally reflect human perception on the shape. By considering local context in the 3D shape, Lee et al. [9] introduce the so-called mesh saliency, defined by averaging the curvature in a local area as its importance. Then the best view is computed as the one observing the largest amount of mesh saliency among a set of sampled views. Noting that previous works are still limited to surface shapes, Takahashi et al. [10] propose a decomposition based best views for volume models. They decompose the entire volume into feature components, and determine the global best view as the weighted average of locally optimal views for the components.

Some other best view definitions by high level semantic analysis on the 3D shapes are also introduced in best view computation. Considering the best views should be as dissimilar from each other as possible, Denton et al. [2] compute the best views by filtering the "equivalent views" from viewing space. Polonsky et al. [11] propose a view descriptor by combing the cues like surface area entropy, visibility ratio, curvature entropy, silhouette length, silhouette entropy, topology complexity, and surface entropy. Then the best view is computed by maximizing the descriptor. Mortara and Spagnuolo [3] define the best view as seeing the most meaningful components of the 3D shape. So they segment 3D shapes into semantic components, and compute the best views by maximizing the visibility of relevant and meaningful components. Laga [12] defines the best view of a 3D object as the one able to discriminate objects from each other in the database. However, this approach needs a database training and then performs best view query. Although previous approaches can provide good views for 3D shapes, it is hard to say these views conform to human perception, whereas in this paper, our approach aims at the faithful visual perception when observing 3D shapes.

Since our approach is based on web images, we briefly review some works in this field. Recent years have witnessed a rise in the research of Internet image based processing, such as photo tourism [13], image completion [14], photo montage [15], and so forth. These techniques use the intrarelated web images to create new images that obey human concepts. In this paper, we will exploit the human intuition on 3D shapes involved in the relevant web images and then compute the best views from the images. To the best of our knowledge, it is the first time to handle the problem in this way.

3 Algorithm overview

The main idea of our algorithm is to extract perceptual view information from the relevant web images. Formally, given a 3D shape M, a series of related web images $\mathcal{I}$ containing (or similar to) M are searched and collected. From the view space $\mathcal{V}$ around M, the best view $\tilde{v} \in \mathcal{V}$ is selected as the one attaching the most web images which are possibly captured from that viewpoint. Our algorithm includes three stages: acquisition of relevant web images, web image voting and viewpoint clustering. Figure 1 shows the work flow of our algorithm and described as below.

In the stage of web image acquisition (Sect. 4), we first search the web images relevant to the 3D shape from the Internet. Then, downloading and filtering of web images are performed to compose a voter database for the 3D shape, which retains only the web images feasible to segment and participates the best view selection. In the stage of web image voting (Sect. 5), each image in the voting database picks the viewpoint with the best matching result. A similarity measurement between 2D image and 3D projection from that viewpoint is employed to conduct the web image votes. Then we record the number of votes for each viewpoint. In the stage of viewpoint clustering (Sect. 6), we perform adaptive clustering on the viewpoints weighted by the number of votes, which results in more robust computation of best view.

4 Acquisition of relevant web images

Given a 3D shape M, we need to search as many relevant images as possible from the Internet, and finally build the web image database $\mathcal{I} = \{I_k\}$. There are two typical

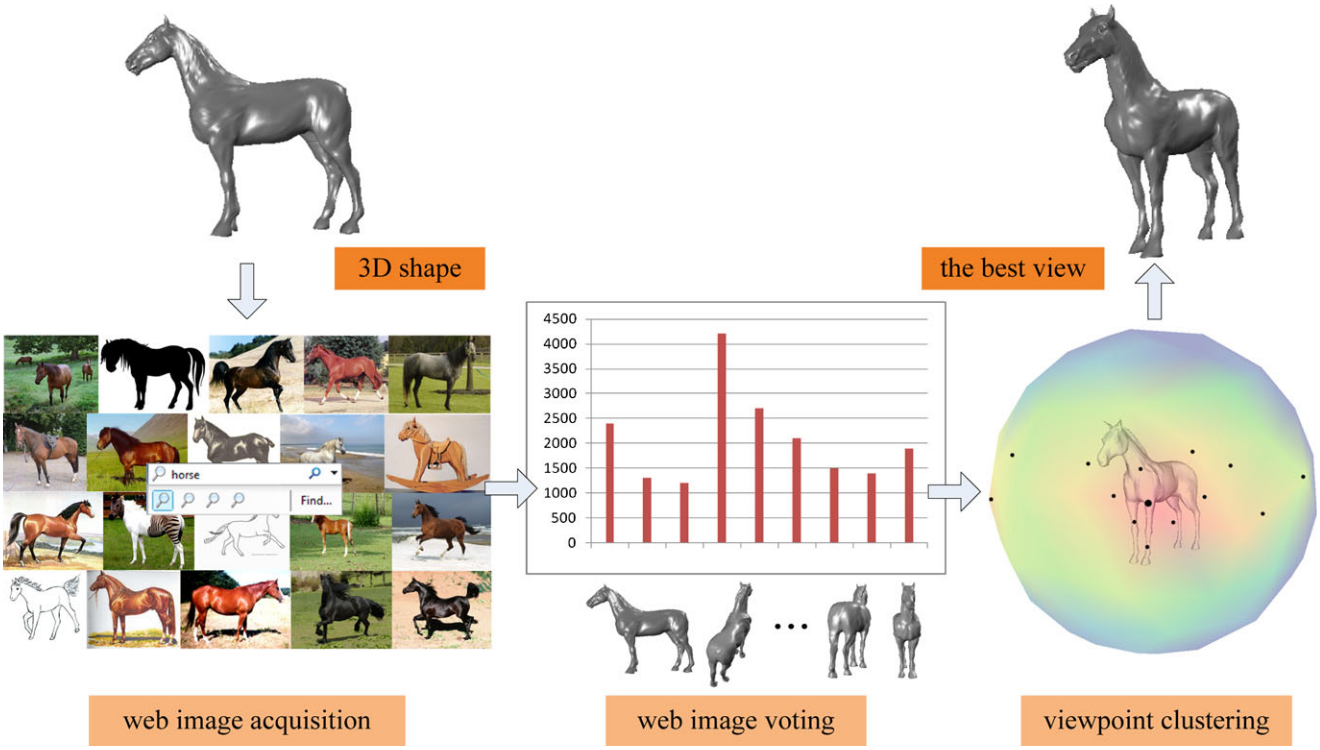

Fig. 1 Overview of our method. Our method has three stages: web image acquisition, web image voting, and viewpoint clustering. The best view is selected to be the viewpoint cluster center with the most votes

approaches for image searching: text-based and content-based [16]. Theoretically, content-based searching can retrieve more accurate images with contents conforming to the 3D shape. But, as it is only in exploration phase and not extended in practical use [17], we use text-based image searching but with next post-processing to obtain the desired images in our application. For the 3D shape M, user inputs a keyword to indicate what M represents. For example in Fig. 2, we can use "car" or "sedan" as keyword to collect images possibly containing a car. Also, to make the searching images more accurate, we should input the keywords as descriptive as possible, which would make the downloaded images more practical.

However, given the massive data of web images, we may get some bad images with irrelevant contents or even with no object corresponding to the 3D shape (see Fig. 2). So, we need to further classify images in $\mathcal{I}$ to filter out the irrelevant images and segment the rest to get the foregrounds that are possibly profiles when observing M from different views. As the background of the images from Internet are different from each other, filtering and segmenting all the images automatically is a challenge task, and here we do not intend to address them seriously. Chen et al. [15] introduce a set of "algorithm-friendly" filters, which are effective for web image classifying. In our approach, we only need to get the foreground of each image, so we use their saliency filter, scene item segmentation, and content consistency filter to process images in $\mathcal{I}$. After processing images with these filters, we obtain the image database that is allowed to vote the views of M. For convenience, we still use $\mathcal{I}$ to denote the web image database after filtering and segmentation, which would be involved in the view election in the next section.

5 Web image voting

As a 3D shape M can be observed from any point with arbitrary orientation, the view space is effectively infinite [11]. We can borrow the measurement of "three-quarter views" and "normal clustering" [11] to generate a set of candidate viewpoints in the computation. Bur for simplicity, we just sample the viewpoints on a sphere surrounding M uniformly. Concretely, we define a bounding sphere $O(o, 2r)$, whose center o is located at the gravity center of M, and its radius r is the half diagonal length of oriented bounding box of M. Then we uniformly sample points from $O(o, 2r)$ to construct a set of candidate views as $\mathcal{V} = \{v_i\}$. As shown in Fig. 3, we can get a corresponding projection P_i of M on the view plane perpendicular the view direction $v_i - o$ from each sampled viewpoint v_i.

After getting the candidate views $\mathcal{V}$, each image $I_k \in \mathcal{I}$ will be appointed to tally a vote to a view v_k from which the image is most possibly captured. In this section, we compute the similarity measurement between image I_k and projection

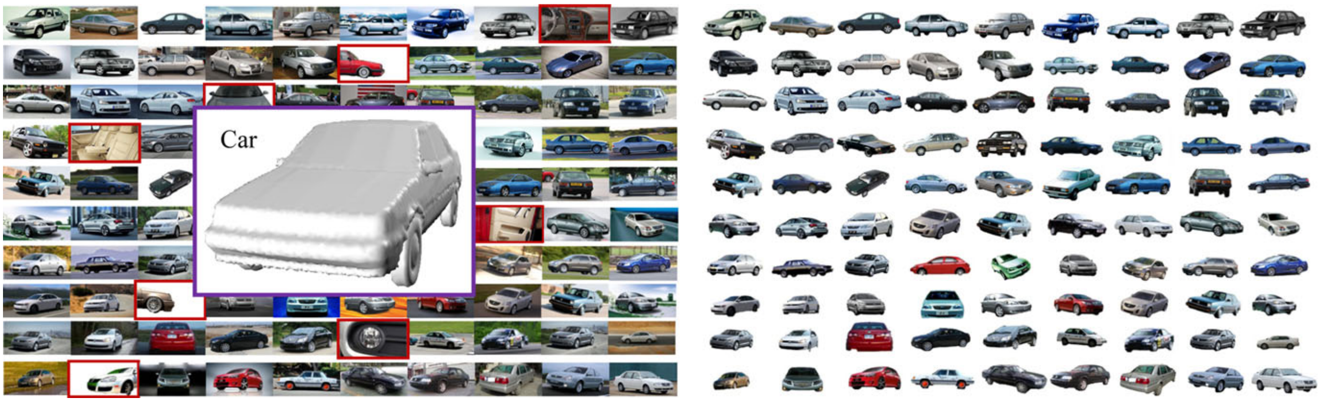

Fig. 2 *Left*: Given a 3D shape and its key word, millions of web images can be obtained from the Internet. “Algorithm-friend” filters are used to exclude irrelevant images (highlighted by *red rectangles*) which are difficult to recognize and segment the contents. *Right*: The rest images are segmented to retrieve the regions relevant to the 3D shape

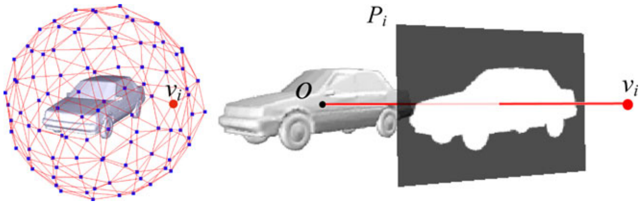

Fig. 3 *Left*: Candidate views sampled on the bounding sphere uniformly. *Right*: For a specified view v_i, 3D shape is projected on the corresponding view plane P_i

P_k to get the similarity between them, and image votes are elected to the view with the best matching. Here, we define the similarity measurement $d_{i,j}$ between I_i and P_j as

$$d_{i,j} = w_1 A(I_i, P_j) + w_2 C(I_i, P_j) + w_3 S(I_i, P_j) \tag{1}$$

where A defines area similarity, C judges silhouette similarity, S measures the saliency disparity between I_i and P_j, and $w_{1,2,3}$ are the weights to control the trade off among the three items. Next, we will introduce how to compute each measurement item in (1).

5.1 Area similarity

For a 3D shape, the visible information may differ dramatically when observing from distinct views. Hence, regional difference between image and 3D projection is an efficient similarity measurement. Formally, area similarity $A_{i,j}$ between I_i and P_j is defined as

$$A_{i,j} = \exp\left(-\frac{\mathrm{Area}(I_i - P_j) + \mathrm{Area}(P_j - I_i)}{2 \cdot \mathrm{Area}(I_i \cap P_j)}\right) \tag{2}$$

where $I_i - P_j$ denotes the region in I_i but not in P_j, $P_j - I_i$ denotes the region in P_j but not in I_i, and $I_i \cap P_j$ is the overlapping region between I_i and P_j.

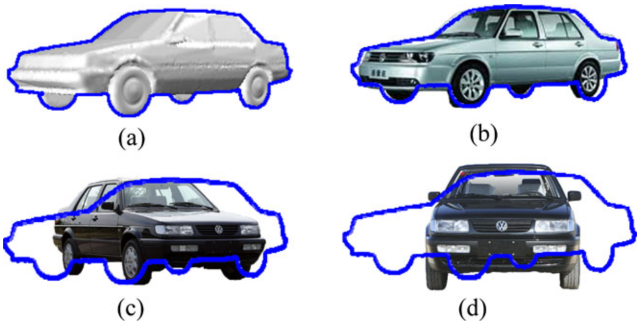

Fig. 4 Area similarity is measured by computing the overlap ratio between projected 3D shape (**a**) and the web images (**b**–**d**) with the canonical transformation. The values of (2) for (**b**–**d**) are 0.98, 0.77 and 0.72

When taking a photo for the same object, the imaging may affected by camera rotation and translation. To overcome this problem, we need to relocate I_i and P_j into a common coordinate system first. In this paper, we obtain the normalized images by applying principle component analysis (PCA) to get the canonical transformation of I_i and P_j. Concretely, the center of mass of I_i and P_j is translated to the origin, and to make sure the larger principle component becomes 1, each principle axes coincides with x–y axes is multiplied with a proper scaling. After performing canonical transformation, we can compute the area similarity according to (2). The computation of area similarity between the profile of 3D shape from some viewpoint and the web images is shown in Fig. 4.

5.2 Silhouette similarity

As an important visual cue for 3D (2D) shape description, silhouette is an effective characteristic to object recognition, especially in the absence of color and texture information [18]. Since silhouette represents the geometrical characteristic of an object, in many cases, only silhouette is sufficient to discriminate different objects from each other. Here in our setting, we want to determine the similarity between the silhouette of image I_i and silhouette of 3D shape M from a specified view v_j.

The silhouette of image region can be effectively detected from its foreground mask (see Sect. 3) by classical edge detection operators, e.g., Canny edge operator, whilst 3D silhouette is dependent on the position of view point. As mentioned by Gooch et al. [19], there are many ways to approach it. In this stage, only external silhouette is needed, so we can first retrieve the image from frame buffer when rendering the 3D shape through projection P_j, then employ Canny operator on the retrieved image to get the silhouette. In sequel, U_i and V_j are used to represent the silhouette of I_i and P_j separately, and the task is to compute the silhouette similarity between U_i and V_j.

Hu [20] introduced moment invariants into visual pattern recognition, and derived seven moment invariants that are invariant under translation, scaling and rotation of the object. In our work, we only choose the first three parameters for simplification. The first three moment invariants are defined as:

$$\begin{aligned} &\varPhi_1 = \upsilon_{20} + \upsilon_{02} \\ &\varPhi_2 = (\upsilon_{20} - \upsilon_{02})^2 + 4{\upsilon_{11}}^2 \qquad (3) \\ &\varPhi_3 = (\upsilon_{30} - 3\upsilon_{12})^2 + (3\upsilon_{21} - \upsilon_{03})^2 \end{aligned}$$

and υ_{pq} is the normalized central moment of order $p+q$, defined as follows:

$$\upsilon_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\omega}}, \quad \omega = \frac{p+q}{2} + 1 \qquad (4)$$

where, μ_{pq} is the central moment of order $p+q$. For a given binary image G, μ_{pq} is defined as:

$$\mu_{pq} = \int\int_G (x - x_t)^p (y - y_t)^q \, dx \, dy \qquad (5)$$

where (x_t, y_t) are the coordinates of the center of gravity of image G.

For the web images in $\mathcal{I}$ and 3D projection P_j, its silhouette can be represented as a 3-dimension vector $e(\cdot) = (\varPhi_1, \varPhi_2, \varPhi_3)$. As shown in Fig. 5, if two silhouettes are similar, they have less deviation between the moment invariants. So, given a silhouette pair (U_i, V_j), their silhouette similarity can be defined as

$$C_{i,j} = \exp(-d_{i,j}) \qquad (6)$$

where $d_{i,j}$ is

$$d_{i,j} = \sum_{k=1}^{3} \big(\varPhi_k(U_i)/\varPhi_k(V_j) - 1\big)^2 \qquad (7)$$

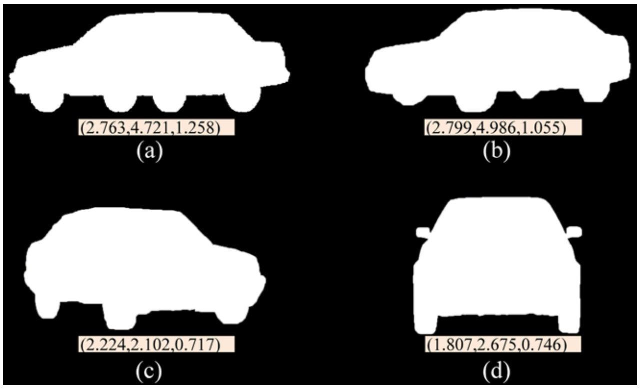

Fig. 5 The vector on the bottom of each silhouette is their moment invariants respectively. The values of (6) for (**b**–**d**) are 0.957, 0.558, and 0.623

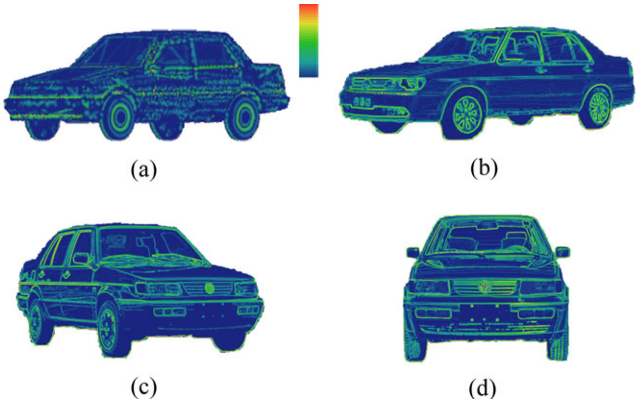

Fig. 6 The saliency of 3D shape (**a**) is computed with mean curvature [22]. While for 2D images (**b**–**d**), saliency is measured by the gradient of contrast distribution in color space [23]. The corresponding 3D shape and web images are shown in Fig. 4. The values of (9) for (**b**–**d**) are 0.0079, 0.00018, 0.0007

Actually, for each web image I_i, (7) measures the moment deviation with respect to the given 3D projection P_j. So, the web image with closer silhouette to the 3D projection would gain smaller deviation.

5.3 Saliency similarity

Area and silhouette are both based on the mask of foreground, which can only represent the geometry information, and analyze the images in a globally cognitive manner. As the best view intends to shift human attention to informative features [4], local salient visual features are also important cues for object recognition. Besides, due to the symmetry of 3D shape, area and silhouette might be exactly the same when observing in the opposite view directions. Hence, we further resort to saliency similarity between the image I_i and the projection P_j.

However, there seems to be no apparent connection between images and 3D shape, as they are obtained in different ways: images are mostly captured by cameras, while

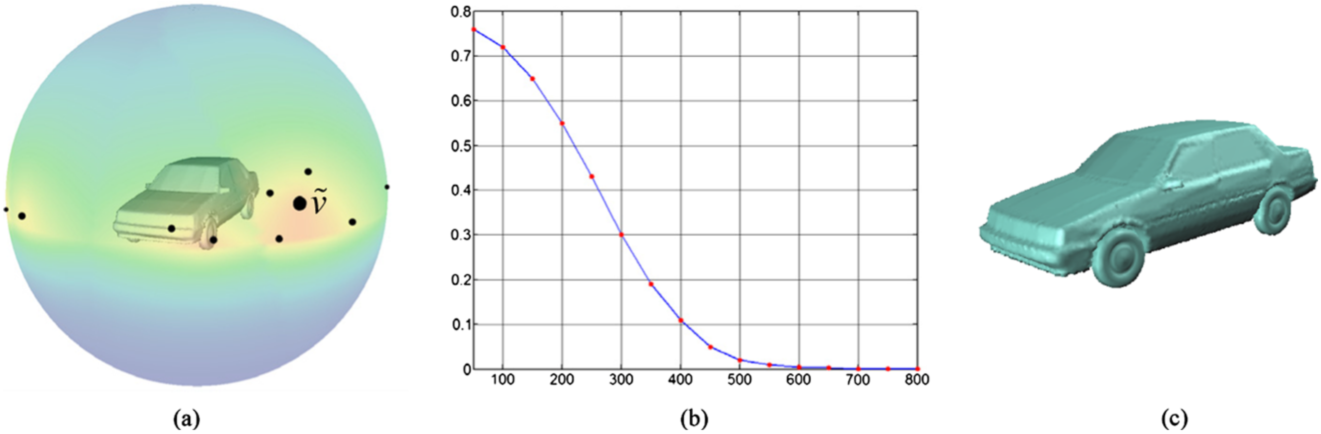

Fig. 7 **a** The best view $\tilde{v}$ is selected from cluster centers (*black points*) of sampled viewpoints with their votes. **b** The location deviation of best viewpoint decreases when the sampling number raises. **c** The 3D shape is observed from the best view $\tilde{v}$

3D shape is recovered by range scanner. Nevertheless, psychologists have investigated that visual attention is indeed attracted by salient stimuli that "pop out" from their surroundings [21], which inspires us to measure saliency similarity using some pop-out type saliency defined from image processing and geometry processing.

For 2D images, there are many approaches to address attention detection [23, 24]. Here, we use the visual attention detection approach based on the color contrast [23]. More specifically, for a pixel x_k with color a_i (R, G, or B channel of the image), the attention value is computed as

$$\mathrm{Att}(x_k) = \sum_{n=0}^{255} f_n D(m, n) \tag{8}$$

where f_n is the frequency of pixel value a_n, and D is the color difference. Att can be seen as the first order derivative of color distribution, which can detect the salient regions with dramatic color difference. To "pop out" these salient regions, we define the image saliency SalI as the gradient of attention values, i.e., $\mathrm{SalI} = \|\nabla \mathrm{Att}\|$. Figure 6(b–d) show the saliency defined by SalI.

For 3D shape, saliency can be locally characterized by a descriptor such as curvature. Curvature seems to draw more attention [9], as curvature is the second order fundamental form and high curvature parts usually possess the anatomical meaning mathematically. In our setting, we employ the mean curvature [22] to represent the saliency of each vertex in the 3D shape. So, for the 3D shape projection P_j, mean curvature $\mathrm{SalM}(x_k)$ is recorded for each viewed point x_k (see Fig. 6(a)). Subsequently, after normalization of SalI and SalM into the range [0, 1], saliency similarity between I_i and P_j can be defined as

$$S_{i,j} = \exp\left(-\sum_{x_k \in I_i \cap P_j} \left\| \mathrm{SalI}(x_k) - \mathrm{SalM}(x_k) \right\|\right) \tag{9}$$

Here, I_i and P_j must be aligned into a common coordinates system by the corresponding canonical transformation as described in Sect. 5.1 before computing (9).

6 Viewpoint clustering

Based on the web image voting, we can summarize the votes number for each candidate view $v_i \in \mathcal{V}$, which naturally reflects human preference observing the 3D shape. Then the best view $\tilde{\mathrm{v}}$ is selected as the one with the most votes.

Obviously, the position of best view is dependent on the viewpoint sampling on the bounding sphere. To avoid distraction caused by sampling, we further use mean-shift clustering [25] to conduct a more stable best view. Concretely, for each sampled viewpoint v_i, we define a 4-dimensional vector $\rho_i = (x_i, y_i, z_i, n_i)$, where (x_i, y_i, z_i) is the position of viewpoint v_i, and n_i is its votes number. After mean-shift clustering, we get a group of clusters as $\mathcal{C} = \{C_1, C_2, \ldots, C_N\}$, where N is the cardinal of clusters, and C_i contains views around the corresponding cluster center c_i (see Fig. 7(a)). By comparing the value of n_i at each cluster center c_i, the best view is defined to be the one with the largest number of votes. Figure 7(b) shows the best view position $\tilde{v}$ converges with the number of sampled viewpoint increasing, because the deviation $\Delta_i = |\tilde{v}_{i+1} - \tilde{v}_i|$ decreases with the increase of the number of sampled viewpoints. Finally, we obtain the best view for the given 3D shape (see Fig. 7(c)).

7 Experiments

In our experiment, we performed our approach on shapes from the Princeton Shape Benchmark (PSB) [26] to select their best view. We sampled 500 candidate views uniformly

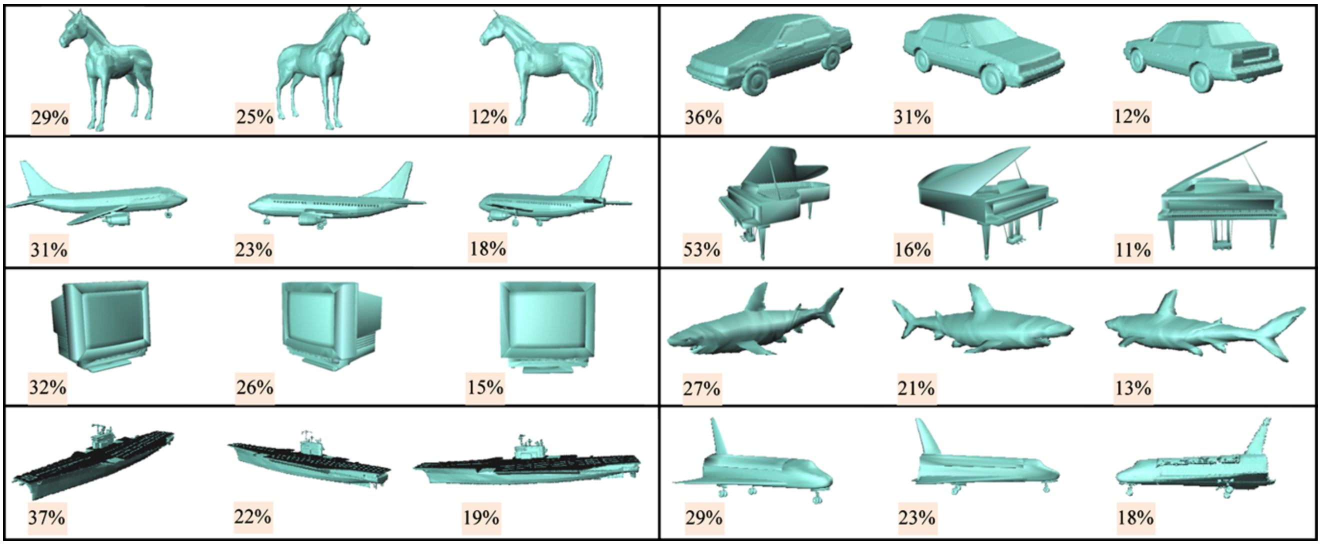

Fig. 8 The best three views selected by our method are shown for each 3D shape. The numbers denote percentage of votes obtained from relevant images

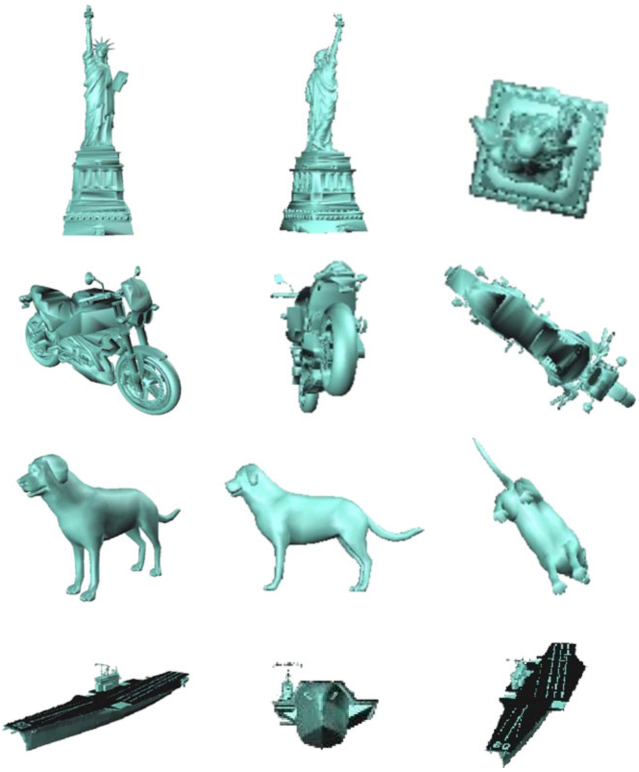

Fig. 9 Comparison of best views obtained by three different approaches: our method (*left*), viewpoint entropy [6] (*middle*) and mesh saliency [9] (*right*)

on the bounding sphere of each 3D shape. We searched about 60,000 images from the Internet for each 3D shape, and about 3000 images survive to the election on best view after image filtering (see Sect. 4). Since the Internet changes everyday, the web images are still increasing at every moment and the best view is allowed to update when new images are discovered and involved in the voting. Our experiments are performed on 2.8 Hz PC with 2G RAM, and most time is consumed in image segmentation (2.5 seconds per image) in the stage of Internet image acquisition, silhouette, and saliency similarity computation (4 seconds per image) in the stage of images voting, and much less time consuming in viewpoint clustering (1.2 seconds).

Some results of our approach is shown in Fig. 8. We list the best three views for each 3D shape and the percentage of votes obtained from the filtered web images. The results shows that people like to observe objects in front but usually with slightly oblique shift, which also agrees with the aesthetic criterion [11]. These views actually obey the human perception on observing them especially in photography. Also, we compare our approach with two other best view selection approaches [6, 9] that use different "goodness" definition (see Fig. 9). Viewpoint entropy [6] selects the best view with maximal visual entropy, which might fail to provide a view that is easy to recognize the 3D shape. Mesh saliency [9] tries to see as many salient features as possible, but ignores the semantic meaning of 3D shape. So, the selected best view might completely disagree with intuition when observing the shape. In contrast, our approach can provide better views that conform to human perception, as we encourage people to select the best view via captures images as well take advantage of people's views shared on the Internet.

User study Best view selection is a much intuitive task for observing 3D shapes, so we conducted a user study to qualitatively evaluate the best view results computed by our approach. We involved 50 common people to rank the best

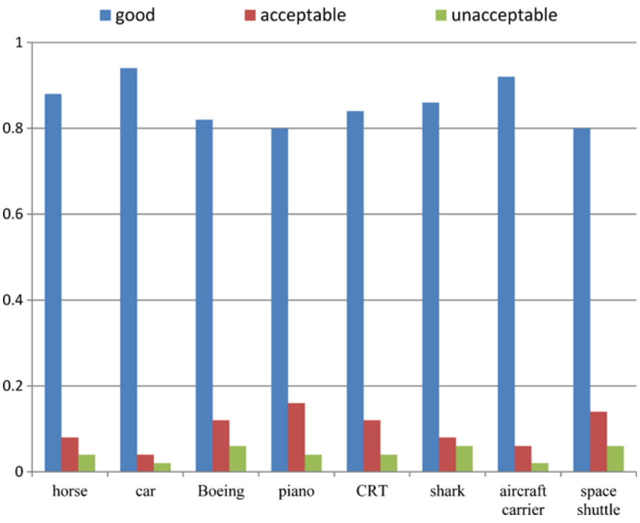

Fig. 10 Statistics of user study

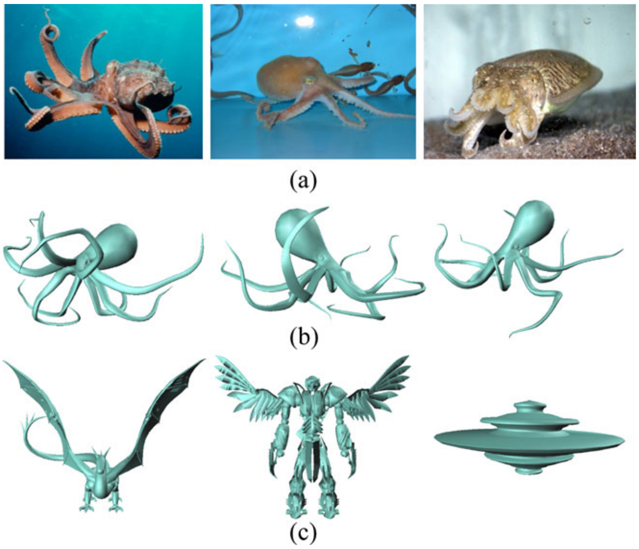

Fig. 11 Failure result by our approach. **a** Web images about octopus. **b** Left to right: best views obtained by our method, viewpoint entropy [6] and mesh saliency [9]. **c** 3D shapes are rare in real world

views for the 3D shapes listed in Fig. 8. We adopted the view criterion in [3] and asked the people to tag the computed best views as "good," "acceptable," or "unacceptable." Figure 10 shows the statistics result of our user study. We recorded the percentage of "good," "acceptable," and "unacceptable" tagging number for each 3D shape ranked by different people. The results are very satisfactory based on the statistics. It can be seen that our approach is able to advise "good" views in high rate, especially for the 3D shapes that are common to be photographed and slightly varied in the real world, e.g., car.

Limitations Our approach favors best view of rigid or quasi-rigid 3D shapes like cars, instruments, fishes, and so on, whose pose with slight shape variance in daily life. However, for the nonrigid shapes, especially for the articulated shapes, our approach might result in unreasonable best views in contrast to human common sense. For example in Fig. 11(a), the tentacles of octopus usually swing with arbitrary motion, which causes the appearance of the octopus to constantly change. Thus, our approach fails to obtain the best view from web images compared with other approaches (see Fig. 11(b)). Moreover, our approach make use of web images which contains the object the 3D shape represents, but as shown in Fig. 11(c), we may face the 3D shapes that are rare in our daily life and retrieve few relevant web images, which disable our approach to best view computation. Additionally, our approach relies on the segmentation of web images to extract the foreground objects for best view election. However, automatic segmentation might generate incorrect foreground, which subsequently causes incredible voting. In this case, we have to apply extra interaction which would add more efforts.

8 Conclusion

We present a novel approach to compute the best view of a 3D shape driven by web images. Our approach encourages human perception directly applied in best view selection. Due to the statistics on the various web images, our approach is able to select semantic and stable views that coincide with the common sense.

As the future work, we need to improve the stage of web image acquiring, both for accuracy and efficiency. The filtered web images can be further weighted with some subjective assessment: If images are captured by professional photographers, the votes they give would be more convincing. Additionally, as the database would contain a huge number of images, the computational efficiency of our approach limits its usefulness, which needs to be disposed by parallel processing or advanced approach such as cloud computing.

Acknowledgements We thank the anonymous reviewers for their helpful comments. This work was partly supported by the Program for New Century Excellent Talents in University (No. NCET-09-0635).

References

1. McMillan, L., Bishop, G.: Plenoptic modeling: an image-based rendering system. In: Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques, pp. 39–46 (1995)
2. Denton, T., Demirci, M.F., Abrahamson, J., Shokoufandeh, A., Dickinson, S.: Selecting canonical views for view-based 3-d object recognition. In: International Conference on Pattern Recognition, vol. 2, pp. 273–276 (2004)
3. Mortara, M., Spagnuolo, M.: Semantics-driven best view of 3d shapes. Comput. Graph. **33**(3), 280–290 (2009)

4. Blanz, V., Tarr, M.J., Bülthoff, H.H., Vetter, T.: What object attributes determine canonical views? Perception **28**(5), 575–600 (1999)
5. Fu, H.B., Cohen-Or, D., Dror, G., Sheffer, A.: Upright orientation of man-made objects. ACM Trans. Graph. **27**(3), 42–48 (2008)
6. Vázquez, P.P., Feixas, M., Sbert, M., Heidrich, W.: Viewpoint selection using viewpoint entropy. In: Proceedings of the Vision Modeling and Visualization Conference, pp. 273–280 (2001)
7. Vázquez, P.P., Feixas, M., Sbert, M., Llobet, A.: Viewpoint entropy: a new tool for obtaining good views of molecules. In: Proceedings of the Symposium on Data Visualisation, pp. 183–188 (2002)
8. Page, D.L., Koschan, A.F., Sukumar, S.R., Roui-Abidi, B., Abidi, M.A.: Shape analysis algorithm based on information theory. In: International Conference on Image Processing, vol. 1, pp. 229–232 (2003)
9. Lee, C.H., Varshney, A., Jacobs, D.W.: Mesh saliency. ACM Trans. Graph. **24**(3), 659–666 (2005)
10. Takahashi, S., Fujishiro, I., Takeshima, Y., Nishita, T.: A feature-driven approach to locating optimal viewpoints for volume visualization. In: IEEE Visualization, pp. 495–502 (2005)
11. Polonsky, O., Patané, G., Biasotti, S., Gotsman, C., Spagnuolo, M.: What's in an image? Vis. Comput. **21**(8), 840–847 (2005)
12. Laga, H.: Semantics-driven approach for automatic selection of best views of 3d shapes. In: Eurographics Workshop on 3D Object Retrieval (2010)
13. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. ACM Trans. Graph. **25**(3), 835–846 (2006)
14. Hays, J., Efros, A.A.: Scene completion using millions of photographs. ACM Trans. Graph. **26**(3) (2007)
15. Chen, T., Cheng, M.M., Tan, P., Shamir, A., Hu, S.M.: Sketch2photo: internet image montage. ACM Trans. Graph. **28**(5), 124–133 (2009)
16. Raimondo, S., Gianluigi, C., Silvia, Z., Istituto, T., Infomatiche, M.: A survey of methods for colour image indexing and retrieval in image databases, pp. 183–211 (2001)
17. Schroff, F., Criminisi, A., Zisserman, A.: Harvesting image databases from the web. In: International Conference on Computer Vision, pp. 1–8 (2007)
18. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. **24**(4), 509–522 (2002)
19. Gooch, B., Hartner, M., Beddes, N.: Silhouette algorithm. In: ACM SIGGRAPH Course Notes (2003)
20. Hu, M.K.: Visual pattern recognition by moment invariants. IRE Trans. Inf. Theory **8**(2), 179–187 (1962)
21. Connor, C., Egeth, H., Yantis, S.: Visual attention: bottom-up versus top-down. Curr. Biol. **14**(19), 850–852 (2004)
22. Meyer, M., Desbrun, M., Schröder, P., Barr, A.H.: Discrete differential geometry operators for triangulated 2-manifolds. Math. Vis. **3**(7), 1–26 (2002)
23. Zhai, Y., Shah, M.: Visual attention detection in video sequences using spatiotemporal cues. In: ACM International Conference on Multimedia, pp. 815–824 (2006)
24. Cheng, M.M., Zhang, G.X., Mitra, N., Huang, X.L., Hu, S.M.: Global contrast based salient region detection. In: Computer Vision and Pattern Recognition (2011)
25. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 603–619 (2002)
26. Shilane, P., Min, P., Kazhdan, M., Funkhouser, T.: The Princeton shape benchmark. In: Proceedings of the Shape Modeling International, pp. 167–178 (2004)

Hong Liu is currently a Ph.D. student in School of Electronic and Information Engineering, Xi'an Jiaotong University, China. His research interests include video and image processing.

Lei Zhang received his B.S. and Ph.D. degrees in Applied Mathematics from Zhejiang University, China, in 2004 and 2009. His research interests include computer graphics, image and video processing.

Hua Huang is currently a Professor in School of Electronic and Information Engineering, Xi'an Jiaotong University, China. He received his B.S. and Ph.D. degrees from Xi'an Jiaotong University, in 1996 and 2006. His main research interests include image and video processing, pattern recognition.