SALIENT VIEW SELECTION BASED ON SPARSE REPRESENTATION

Yi-Chen Chen[1]*, Vishal M. Patel*[1]*, Rama Chellappa*[1]*, and P. Jonathon Phillips*[2]

[1]Department of Electrical and Computer Engineering and the
Center for Automation Research, UMIACS, University of Maryland, College Park, MD
[2]National Institute of Standards and Technology, Gaithersburg, MD
{chenyc08, pvishalm, rama}@umiacs.umd.edu   jonathon.phillips@nist.gov

ABSTRACT

A sparse representation-based approach is proposed to find the salient views of 3D objects. Under the assumption that a meaningful object can appear in several perceptible views, we build the object's approximate convex shape that exhibits these apparent views. The salient views are categorized into two groups. The first are *boundary representative views* that have several visible sides and object surfaces attractive to human perceivers. The second category are *side representative views* that best represent views from sides of the approximating convex shape. The side representative views are class-specific views that possess the most representative power compared to other within-class views. Using the concept of *characteristic view* class, we first present a sparse representation-based approach for estimating the boundary representative views. With the estimated boundaries, we determine the side representative view(s) based on a minimum reconstruction error.

Index Terms— Salient view, characteristic view class, view geometry, sparse representation, compressive sensing.

1. INTRODUCTION

The concept of characteristic views was first proposed in [1][2] for object recognition. It was defined in such a way that two views belonging to the same Characteristic View Class (CVC) are topologically equivalent, and they can be related by a 3D transformation which consists of geometric rotation, translation and perspective projection [3]. [3] proposes a framework to partition the viewing space and to find the set of characteristic views for planar-faced solid objects. This work was later extended in [4], which essentially computes the characteristic views of objects with curved-surface.

There are a number of approaches for describing what is contained in a view [5], [6]. For view-based representations, human perceivers are influenced by factors such as the familiarity with the object being viewed, the similarity of a given view to known views of visually-similar objects and the pose of the object [5]. The three-quarter views with all visible front, top and side, are often used as candidate views. As noted in [7], these are essentially the views that most humans prefer when looking at an object. These views are also known as the *canonical views* [6].

In [8], saliency was defined as the amount of energy not captured by the basis set in an eigenspace representation. A greedy algorithm was proposed for subset selection where the saliency of every ensemble view is first computed and then the view with the highest saliency is added to the subset. The subset is then updated using the eigenspace representation updating algorithm [9][10] so that the task of salient view selection can be realized in a dynamic environment.

In recent years, Sparse Representations (SR) have emerged as a powerful tool for efficient processing of data in non-traditional ways. Motivated by the success of SR in many computer vision and image processing applications [11], [12], we propose an SR-based approach to select the salient views of an object [6], [7]. Given an object, we first find its approximate convex shape which exhibits apparent sides. A side view class (SVC) is defined as a set of views of the corresponding side of the shape, while a boundary view class (BVC) refers to views where two or more sides can be seen simultaneously. Fig. 1(a) illustrates distinct regions of SVCs and BVCs using an approximate convex shape of a given object. The shape in this figure consists of four sides, which give four SVCs and four BVCs under orthographic projection. These eight classes are exactly the eight CVCs of the approximate convex shape. With the concept of object's approximate convex shape and its sides, we categorize salient views into two categories: *boundary representative views* (BRVs) which have more visible sides and object surfaces, and therefore are more attractive from a human perception point of view; and *side representative views* (SRVs) which best describe the underlying SVCs. In Fig. 1(a), BRVs and SRVs are views seen from directions marked with red and blue arrows, respectively. Fig. 1(b) shows the block diagram of the proposed two-stage approach to find salient views. Views are extracted from a video sequence, cropped and properly resized. In the first stage (in blue) the boundary scores are computed using a spread metric and BRVs are estimated. Next the SVCs are determined. In the second stage (in green), the SRV(s) are chosen that best represent the associated side by minimizing a representation error.

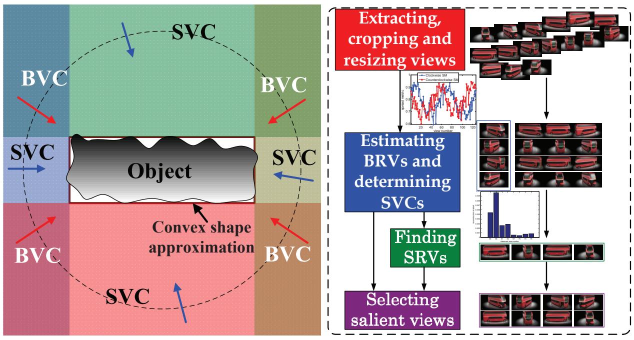
This paper is organized as follows. Section 2 presents a way to estimate the BRVs of an object. In Section 3, we describe our approach to determine the SRVs. Experiment results on three 3D-view example sequences are presented in Section 4. Section 5 concludes the paper with a brief summary and discussions.

2. ESTIMATING BOUNDARY REPRESENTATIVE VIEWS

For any convex object, the type-A planes described in [3] are sufficient to partition the space into CVCs. Whenever two views belong to the same CVC, every viewable point in one view is also viewable in the other view, and vice versa. Following this idea and using the approximate convex shape of a given object, we present an approach to estimate BRVs as follows.

Spread Metric: Let $\mathbf{S}_m$, where $m \in \{1, 2, ..., N\}$, be a subset of the m-th SVC. $\mathbf{S}_m$ consists of a finite number of views that are ap-

This work was partially supported by an ONR grant N00014-12-1-0124.


SVC
BVC
BVC
SVC
Object
SVC
Convex shape approximation
BVC
BVC
SVC
Extracting, cropping and resizing views
Estimating BRVs and determining SVCs
Finding SRVs
Selecting salient views

(a) (b)

Fig. 1. (a) Convex shape approximation and the associated SVC/BVC regions. (b) Block diagram of the proposed salient view selecting approach.

proximately topologically equivalent as they can be related by 3-D transformations. We can further subdivide $\mathbf{S}_m$ into exclusive subgroups $\mathbf{s}_i$'s such that $\mathbf{S}_m = \mathbf{s}_1 \bigcup \mathbf{s}_2 \bigcup ... \bigcup \mathbf{s}_{k_m}$. $\mathbf{s}_i$ contains views that are fairly close to each other as they are viewed from locations with small rotation or translation differences. We further assume $\mathbf{S}_m$ is sufficiently large such that the representation of a view in the class associated with $\mathbf{S}_m$ is sparse under $\mathbf{S}_m$. We use a spread metric denoted by $\mathrm{SM} = 1 - \mathrm{SCI}$, to represent the saliency of a candidate view $\mathbf{z}_j$ to $\mathbf{S}_m$, where SCI stands for Sparsity Concentration Index defined in [13]. SCI is a measure of sparsity of the coefficient representation of a vector under some basis. Low values of SCI (i.e., high SM) indicate that the given view is fairly informative relative to the existing group. Thus, we have

$$\mathrm{SM}_m(\mathbf{z}_j) = \frac{k_m \left(1 - \max\limits_{i \in \{1,2,\ldots,k_m\}} \frac{\|\delta_{m,i}(\mathbf{x}_{m,j})\|_1}{\|\mathbf{x}_{m,j}\|_1}\right)}{k_m - 1}, \quad (1)$$

where $\mathbf{x}_{m,j}$ is the representation of the candidate view $\mathbf{z}_j$ under $\mathbf{S}_m$. That is,

$$\mathbf{x}_{m,j} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{z}_j = \mathbf{S}_m \mathbf{x}. \quad (2)$$

In (1), $\delta_{m,i}(\mathbf{x}_{m,j})$ is a vector whose only nonzero entries are the entries of $\mathbf{x}_{m,j}$ that belong to the i-th subgroup of $\mathbf{S}_m$. It can be shown that $\mathrm{SM}_m(\mathbf{z}_j) \in [0, 1]$. The larger the $\mathrm{SM}_m(\mathbf{z}_j)$ is, the larger the saliency possessed by $\mathbf{z}_j$ relative to $\mathbf{S}_m$. Large $\mathrm{SM}_m(\mathbf{z}_j)$ is a strong indication that $\mathbf{z}_j$ belongs to a subset different from $\mathbf{S}_m$.

Finding boundary representative views: In this section, we describe our method for finding the BRVs. For simplicity, we consider only the 3-D views of an object w.r.t. Y axis rotation ($0° \sim 360°$) under the orthographic projection. Without loss of generality, assume $\{\mathbf{z}_j\}_{j=0}^{K-1}$ are the original full 3-D views of a given object in the clockwise positive direction (i.e., as j increases, it goes in the clockwise direction [1]). After $\mathbf{z}_{K-1}$, the sequence rounds back to $\mathbf{z}_0$ as these are rotated views w.r.t. Y axis. Now, for any given $\mathbf{z}_j$, we calculate its boundary score as follows:

$$\widetilde{\mathrm{SM}}_{W_{j(\beta,\gamma)}}(\mathbf{z}_{j(\alpha)}) = \frac{\gamma \left(1 - \max\limits_{i \in \{1,2,\ldots,\gamma\}} \frac{\|\delta_{W,i}(\mathbf{x}_{W,j(\alpha)})\|_1}{\|\mathbf{x}_{W,j(\alpha)}\|_1}\right)}{\gamma - 1}, \quad (3)$$

where $j(\alpha) \triangleq \mathrm{mod}(j + \alpha, K)$, and

$$W_{j(\beta,\gamma)} \triangleq \left(\mathbf{z}_{j(-\beta-\gamma+1)} \; \mathbf{z}_{j(-\beta-\gamma+2)} \; ... \; \mathbf{z}_{j(-\beta)}\right). \quad (4)$$

In (3), $\mathbf{x}_{W,j(\alpha)}$ is the representation of the candidate view $\mathbf{z}_{j(\alpha)}$ under $W_{j(\beta,\gamma)}$. That is,

$$\mathbf{x}_{W,j(\alpha)} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{z}_{j(\alpha)} = W_{j(\beta,\gamma)} \mathbf{x}. \quad (5)$$

Similar to $\delta_{m,i}(\mathbf{x}_{m,j})$ in (1), here $\delta_{W,i}(\mathbf{x}_{W,j(\alpha)})$ is a masked version of $\mathbf{x}_{W,j(\alpha)}$ such that its only nonzero entry is the one that corresponds to the i-th column vector of $W_{j(\beta,\gamma)}$. In this setting, for a given $\mathbf{z}_j$, we calculate the SM of the view ahead of it by α units of indices, with respect to the set formed from the $(\beta + \gamma - 1)$-th view up to the β-th view behind $\mathbf{z}_j$. That is, this set is formed according to a β-index logged window with size γ.

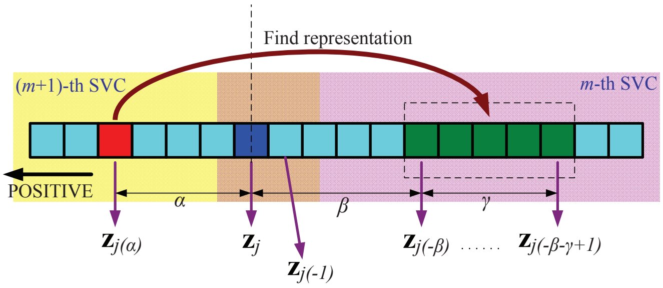


Fig. 2. An illustration of finding the boundary score.

Figure 2 depicts an illustration of how we compute the boundary score. Consider two SVCs: m-th SVC (in color purple) and $(m+1)$-th SVC (in color yellow), and one BVC in between. Since in the beginning no information on SVCs is provided, the choice of basis is unknown and no spread metric can be calculated. Instead, we use a sliding window with a predetermined size γ to select views and form $W_{j(\beta,\gamma)}$ (consisting of views in color green). To find the boundary score at $\mathbf{z}_j$, we calculate the spread metric of $\mathbf{z}_{j(\alpha)}$ which leads $\mathbf{z}_j$ by α units of views, with respect to $W_{j(\beta,\gamma)}$ which lags $\mathbf{z}_j$ by β units of views. Note that α and β should be properly tuned according to not only the complexity of object but also the view sampling interval. If α and β are too small, the spread metric is not obvious as $\mathbf{z}_{j(\alpha)}$ is close to a member of $W_{j(\beta,\gamma)}$. On the other hand, whenever α and β are too large, so are the spread metric since $\mathbf{z}_{j(\alpha)}$ is close to none of $W_{j(\beta,\gamma)}$. In both above cases the spread metric can no longer be a discriminative measure for BRVs. With properly chosen α and β, one could expect the boundary score at $\mathbf{z}_j$ when $\mathbf{z}_j$ is in the BVC (i.e., overlapped region) to be higher than those when $\mathbf{z}_{j(\alpha)}$ and members in $W_{j(\beta,\gamma)}$ are in the same SVC.

3. SIDE REPRESENTATIVE VIEW(S) SELECTION

Representative views can either be interpreted as a sparse representation (i.e., coefficients) under some basis, or can be used as sparse observation where sparse coefficients under some basis can be found. In this section, with representative views regarded as sparse observations, we propose a procedure for finding an object-dependent basis set. We assume that camera parameters are not known.

We assume distinct SVCs are independent of each other. Without loss of generality, we consider the first SVC, $[\mathbf{z}_0 \; \mathbf{z}_1 \; ... \; \mathbf{z}_{k_1-1}]$. Its singular value decomposition (SVD) is $[\mathbf{z}_0 \; \mathbf{z}_1 \; ... \; \mathbf{z}_{k_1-1}] = \mathbf{V}\Sigma\mathbf{U}^T$, where $\mathbf{V} = [\mathbf{v}_1 \; \mathbf{v}_2 \; ... \; \mathbf{v}_L]$ is an L-by-L matrix (L is total number of pixels of an image); $\mathbf{U} = [\mathbf{u}_1 \; \mathbf{u}_2 \; ... \; \mathbf{u}_{k_1}]$ is a k_1-by-k_1 matrix; and

[1] The same analysis follows for the counterclockwise position assumption as well.

$\sigma_1, ...\sigma_{k_1}$ are the eigenvalues (i.e., first k_1 diagonal entries of Σ). It can be shown that

$$\begin{aligned}\mathbf{y}_1 \triangleq \begin{pmatrix} \mathbf{z}_0 \\ \vdots \\ \mathbf{z}_{k_1-1} \end{pmatrix} &= \begin{pmatrix} | & | & \dots & | \\ \mathbf{c}_1 & \mathbf{c}_2 & \dots & \mathbf{c}_{k_1} \\ | & | & \dots & | \end{pmatrix} \begin{pmatrix} \sigma_1 \\ \vdots \\ \sigma_{k_1} \end{pmatrix} \\ &= \begin{pmatrix} — & \mathbf{Q}_1 & — \\ \_ & \vdots & \_ \\ — & \mathbf{Q}_{k_1} & — \end{pmatrix} \begin{pmatrix} \sigma_1 \\ \vdots \\ \sigma_{k_1} \end{pmatrix} = \mathbf{Rw},\end{aligned} \tag{6}$$

where $\mathbf{c}_i$ ($i \in \{1, ..., k_1\}$) is the column-vectorized form of matrix $\mathbf{v}_i\mathbf{u}_i^T$, and each $\mathbf{Q}_j$ ($j \in \{1, ..., k_1\}$) is a L-by-k_1 matrix. Note that $\mathbf{Q}_j$ is not a 1-by-k_1 row vector. In (6), matrix $\mathbf{R}$ is an object-dependent basis set, and $\mathbf{w}$ contains eigenvalues $\sigma_1, ...\sigma_{k_1}$ as coefficients.

Our objective is to select l_1 out of k_1 views as representative views that best represent the SVC, where $l_1 < k_1$. There are $\binom{k_1}{l_1}$ possible selecting ways. Consider one way in which the selected views are $\mathbf{z}_{s_1}, ..., \mathbf{z}_{s_{l_1}}$, which form a column vector $\mathbf{y}_s$. Next, we pick the corresponding matrices $\mathbf{Q}_{s_1}, ..., \mathbf{Q}_{s_{l_1}}$, to form a sub-basis $\mathbf{R}_s$:

$$\mathbf{y}_s \triangleq \begin{pmatrix} \mathbf{z}_{s_1} \\ \vdots \\ \mathbf{z}_{s_{l_1}} \end{pmatrix}; \qquad \mathbf{R}_s \triangleq \begin{pmatrix} \mathbf{Q}_{s_1} \\ \vdots \\ \mathbf{Q}_{s_{l_1}} \end{pmatrix}. \tag{7}$$

We solve the following equation using the ℓ_1 norm:

$$\hat{\mathbf{x}}_{(\mathbf{y}_s)} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{y}_s = \mathbf{R}_s\mathbf{x}. \tag{8}$$

Since $l_1 < k_1$, less constraints are involved in solving (8) than in (6), and one would expect $\hat{\mathbf{x}}_{(\mathbf{y}_s)}$ to be sparser than $\mathbf{w}$. Among all possible $\binom{k_1}{l_1}$ ways, the one which gives the least sparse-to-full reconstruction residual is chosen. In other words, we seek

$$\hat{\mathbf{y}_s} = \arg\min_{\mathbf{y}_s} \|\mathbf{y}_1 - \mathbf{R}\hat{\mathbf{x}}_{(\mathbf{y}_s)}\|_2. \tag{9}$$

The corresponding best reconstruction is closest to $\mathbf{y}_1$, and can be thought of as the one directly reconstructed using sparse observations from these l_1 representative views. It has sparse representation $\hat{\mathbf{x}}_{(\hat{\mathbf{y}_s})}$ under the basis $\mathbf{R}$ defined in (6).

4. EXPERIMENTAL RESULTS

We selected three available sequences of 3-D videos for our experiments: the BUS sequence [2], the HEAD sequence[3] and the JONES sequence [4]. A given video is converted into a set of images, each of which is one view of the object at some particular rotation angle w.r.t. Y axis, ranging from 0° to 360°. Images are cropped and resized in the preprocessing stage. Figure 3 shows these sequences of images. There are 126 views (2.85° increment per view), 32 views ($\sim$ 11.25° increment per view), and 51 views ($\sim$ 7.05° increment per view) for BUS, HEAD and JONES sequences, respectively. In these figures, the sequence of images going from the left to the right in each row, and then from the top row to the bottom row, corresponds to the (camera) clockwise direction. We calculate the spread metrics with $W_{\beta,\gamma}$ sliding in both clockwise (positive) direction, and counterclockwise (negative) direction.

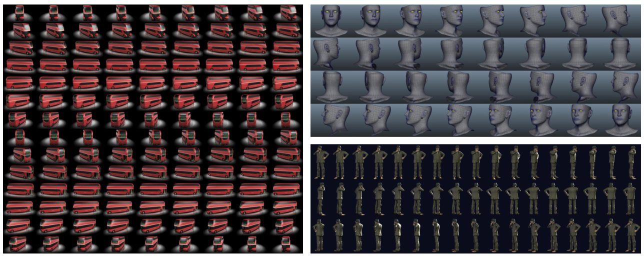
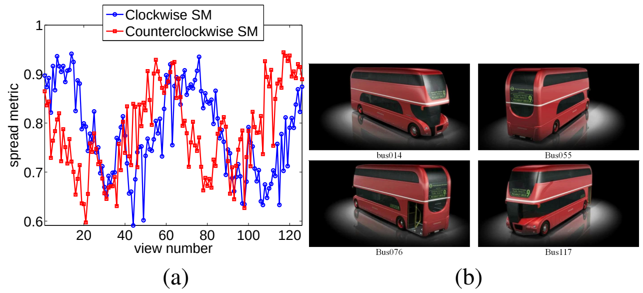


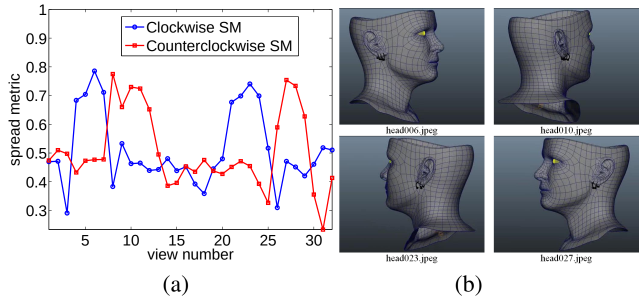
Fig. 3. Sequences of 3-D views. Left: the BUS sequence (126 views); right top: the HEAD sequence (32 views); right bottom: the JONES sequence (51 views).



(a) (b)

Fig. 4. Finding BRVs for the BUS sequence: (a) Clockwise SM and counterclockwise SM. (b) Estimated BRVs.

By assuming that the approximate convex shape has four perceptible sides for the object in each of these sequences, we pick four peaks from spread metric scores. In addition, we use the fact that any two peaks shall be separated by a certain gap, otherwise peaks may be located within the same BVC (the gap is 22.5° for the BUS sequence, and 30° for HEAD and JONES sequences). Figures 4, 5 and 6 show the results. For the BUS sequence, Figure 4(a) suggests that the views with number 014, 055, 076 and 117 are selected as BRVs as shown in Figure 4(b). Likewise, Figures 5(a) and 6(a) suggest views with number 006, 010, 023 and 027, and views with number 010, 022, 035 and 046 as BRVs, shown in Figure 5(b) and 6(b). It is expected that these BRVs are those with more sides and visible surfaces as suggested in [6], [7], and hence human perceivers are more sensitive to them.



(a) (b)

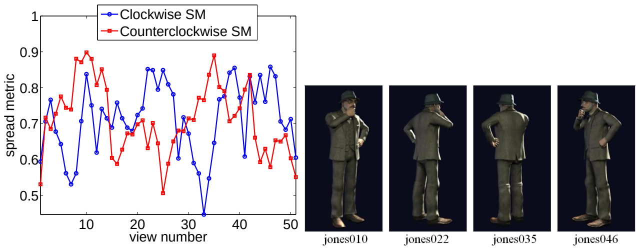
Fig. 5. Finding BRVs for the HEAD sequence: (a) Clockwise SM and counterclockwise SM. (b) Estimated BRVs.

Fig. 7 shows the four SVCs which are separated using the estimated BRVs. Taking into account the overall computation load, we evenly down-sample views in each class, such that each class has no greater than nine views. In Figure 7, we use green lines to mark

[2] http://vimeo.com/3066167

[3] http://vimeo.com/15198240

[4] http://www.youtube.com/watch?v=vq1UeTW6uKE



(a) (b)

Fig. 6. Finding BRVs for the JONES sequence: (a) Clockwise SM and counterclockwise SM. (b) Estimated BRVs.

distinct SVCs. It can be seen that for most cases, views belonging to the same SVC come with more similar poses than those of views that are from distinct SVCs. Figure 8 shows the resulting SRVs. For each SVC, we pick only one view with the minimum sparse-to-full reconstruction error (i.e., $l_1 = 1$). The results of the BUS sequence are shown in Figure 8(a), where views with numbers 034, 070, 096 and 126 are obtained with the minimum residuals calculated by (9) and are representatives of SVCs shown in the first row up to the fourth row at the left top of Figure 7, respectively. Similarly, for the HEAD sequence, views in Figure 8(b) with numbers 009, 014, 027 and 031 are obtained as SRVs of the left bottom 4 rows in Figure 7, whereas views in Figure 8(c) with numbers 016, 030, 039 and 003 are SRVs of those 4 rows of SVCs shown at the right of Figure 7, for the JONES sequence.

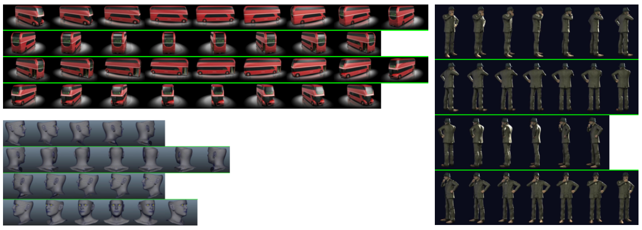
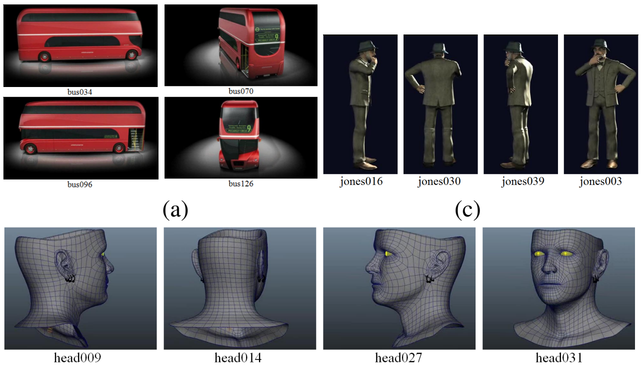


Fig. 7. Estimated 4 SVCs with down-sampled views. Left top: the BUS sequence; left bottom: the HEAD sequence; right: the JONES sequence.

Intuitively, one would expect a SRV to be the side view that capture the most energy compared to other within-class views, and thus have minimum sparse-to-full reconstruction residuals according to (8) and (9). It is not hard to see this phenomenon by comparing representative views in Figure 8 with their associated classes in Figure 7. For all these sequences, the SRVs are generally pretty close to side views: frontal view, left-side view, right-side view and back view. Finally, the salient views are selected from both BRVs and SRVs.

5. CONCLUSION

We presented a two-stage approach based on sparse representation to find salient views of an object. The first stage computes the spread metric and boundary scores to estimate boundary representative views. Using these estimated representative views, full views are roughly partitioned into different side view classes. In the second stage, side representative views are determined that have minimum class sparse-to-full reconstruction residuals. We are currently evaluating the robustness of our approach to noise and occlusions.



(a) (c)

(b)

Fig. 8. SRVs of (a) the BUS sequence (b) the HEAD sequence (c) the JONES sequence.

6. REFERENCES

[1] I. Chakravarty and Herbert Freeman, "Characteristic views as a basis for three-dimensional object recognition," *Proceedings of SPIE*, vol. 336, pp. 37–45, 1982.

[2] Herbert Freeman and I. Chakravarty, "The use of characteristic views in the recognition of three dimensional objects," *Pattern Recognition in Practice*, 1980.

[3] R. Wang and Herbert Freeman, "Object recognition based on characteristic view classes," *Proceedings of IEEE International Conference on Pattern Recognition*, pp. 8–12, 1990.

[4] Shuang Chen and Herbert Freeman, "Characteristic-view modeling of curved-surface solids," *International Journal of Pattern Recognition and Artificial Intelligence - IJPRAI*, vol. 10, pp. 537–560, 1996.

[5] Michael J. Tarr and David J. Kriegman, "What defines a view?," *Vision Research*, vol. 41, pp. 1981–2004, 2001.

[6] Volker Blanz, Michael J. Tarr, and Heinrich H. Bülthoff, "What object attributes determine canonical views?," *Perception*, vol. 28, pp. 575–599, 1999.

[7] Oleg Polonsky, Giuseppe Patané, Biasotti Silvia, Craig Gotsman, and Michela Spagnuolo, "What's in an image? towards the computation of the "Best" view of an object," *The Visual Computer*, vol. 21(8-10), pp. 840–847, 2005.

[8] Y. F., Winkeler, B. S. Manjunath, and S. Chandrasekaran, "Subset selection for active object recognition," *IEEE CVPR*, vol. 2, pp. 511–516, 1999.

[9] S. Chandrasekaran, B. S. Manjunath, Y. F. Wang, J. Winkeler, and H. Zhang, "An eigenspace update algorithm for image analysis," *Graphical Models and Image Processing*, vol. 59, no. 5, pp. 321–332, 1997.

[10] B. S. Manjunath, S Chandrasekaran, and Y. F. Wang, "An eigenspace update algorithm for image analysis," *Proceedings of International Symposium on Computer Vision - ISCV*, pp. 551–556, 1995.

[11] J. Wright, Yi Ma, J. Mairal, G. Sapiro, T.S. Huang, and Shuicheng Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031 –1044, june 2010.

[12] M. Elad, M.A.T. Figueiredo, and Yi Ma, "On the role of sparse and redundant representations in image processing," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 972 –982, june 2010.

[13] John Wright, Allen Y. Yang, Arvinda Ganesh, S. Shankar Sastry, and Yi Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 210–227, 2009.