Iterative Projection and Matching: Finding Structure-preserving Representatives and Its Application to Computer Vision: Supplementary Material

Alireza Zaeemzadeh*, Mohsen Joneidi*, Nazanin Rahnavard, and Mubarak Shah
University of Central Florida
{zaeemzadeh, joneidi, nazanin}@eecs.ucf.edu, shah@crcv.ucf.edu

The supplementary material provided in this document is organized as follows. In Section 1, we present the proofs for the lemmas and propositions stated in the main manuscript. In Section 2 our theoretical results are validated by performing experiments on real and synthetic data. Finally, in Section 3, further experiments are provided to investigate the performance of the proposed approach on several different real datasets.

1. Proofs

Proof of Lemma 1: The inner product between data points of $\boldsymbol{a}_m$'s and $\boldsymbol{v}$ can be calculated as the elements of $\boldsymbol{Av}$. Taking the SVD of $\boldsymbol{A}$, we have $A = \boldsymbol{U\Sigma V}^T$ and

$$|\boldsymbol{Av}| = |\boldsymbol{U\Sigma V}^T\boldsymbol{v}| = \sigma_1|\boldsymbol{u}|.$$

Here, $|.|$ means the element-wise absolute value operation. Since $\|\boldsymbol{u}\|_2 = 1$, there exist at least one element in $\boldsymbol{u}$ (denote as u_i) for which $|u_i| \geq \frac{1}{\sqrt{M}}$. Therefore,

$$|\boldsymbol{v}^T\boldsymbol{a}_i| \geq \frac{\sigma_1}{\sqrt{M}}. \tag{1}$$

This clearly leads to

$$\max_m |\boldsymbol{v}^T\boldsymbol{a}_m| \geq \frac{\sigma_1}{\sqrt{M}}. \tag{2}$$

Proof of Proposition 1:

According to Lemma 1, there exists a data point i for which (1) holds. Given that $\boldsymbol{v}$ has unit length and the rows of $\boldsymbol{A}$ are normalized we have

$$|\rho_i| = \frac{|\boldsymbol{v}^T\boldsymbol{a}_i|}{||\boldsymbol{v}||_2||\boldsymbol{a}_i||_2} \geq \frac{\sigma_1}{\sqrt{M}} = \frac{\sigma_1}{\|\boldsymbol{A}\|_F} = ROM(\boldsymbol{A}),$$

where ρ_i is the correlation between two vectors $\boldsymbol{v}$ and $\boldsymbol{a_i}$ and $\|\boldsymbol{A}\|_F = \sqrt{\sum_{j=1}^{M}||\boldsymbol{a}_j^T||_2^2} = \sqrt{M}$. Accordingly,

$$\max_m |\rho_m| \geq ROM(\boldsymbol{A}). \tag{3}$$

*indicates shared first authorship.

Proof of Lemma 2:

First we compute the derivative of i^{th} eigenvector in terms of matrix $\boldsymbol{C}$ [5],

$$\partial\boldsymbol{v}_i = (\sigma_i^2 I - C)^+\partial\boldsymbol{C}\boldsymbol{v}_i, \tag{4}$$

where $(.)^+$ indicates the MoorePenrose inverse operator. Matrix $\sigma_i^2 I - C$ is singular and its MoorePenrose inverse can be written as follows,

$$(\sigma_i^2 I - C)^+ = \boldsymbol{V\Sigma}_i\boldsymbol{V}^T,$$

where, diagonal elements of $\boldsymbol{\Sigma}_i$ is equal to $1/(\boldsymbol{\lambda} - \lambda_i)$ except the i^{th} diagonal element which is equal to 0. Vector $\boldsymbol{\lambda}$ includes the eigenvalues of $\boldsymbol{C}$. Taking ℓ_2 norm from both side of (4) we have

$$\begin{aligned}\|\partial\boldsymbol{v}_i\|_2 &= \|\boldsymbol{V\Sigma}_i\boldsymbol{V}^T\partial\boldsymbol{C}\boldsymbol{v}_i\|_2 \\ &\leq \|\Sigma_i\|_F\ \|\partial\boldsymbol{C}\|_F = \sqrt{\sum_{j\neq i}\frac{1}{(\lambda_i-\lambda_j)^2}}\|\partial C\|_F.\end{aligned}$$

Note that $\boldsymbol{V}$ is unitary and $\boldsymbol{v}_i$ is normalized and $\sigma_i^2 = \lambda_i$.

Proof of Proposition 2: Obviously $\lambda_1 > \lambda_2$ implies that $s_1 < s_2$. Let us write the expansion of s_1 and s_i for $i > 2$.

$$\begin{aligned}s_1 =& \frac{1}{(\lambda_1-\lambda_2)^2} + \frac{1}{(\lambda_1-\lambda_3)^2} + \cdots \\ & \frac{1}{(\lambda_1-\lambda_i)^2} + \cdots + \frac{1}{(\lambda_1-\lambda_N)^2}\end{aligned}$$

$$s_i = \frac{1}{(\lambda_i-\lambda_1)^2} + \frac{1}{(\lambda_i-\lambda_2)^2} + \cdots + \frac{1}{(\lambda_i-\lambda_N)^2}$$

The $(i-1)^{\text{th}}$ term of s_1 is equal to the first term of s_i. As eigenvalues are sorted in descending order, i^{th} to N^{th} terms of s_1 are less than i^{th} to N^{th} terms of s_i, correspondingly. Thus it is sufficient to show that,

$$\sum_{j=2}^{i-1}\frac{1}{(\lambda_1-\lambda_j)^2} < \sum_{j=2}^{i-1}\frac{1}{(\lambda_j-\lambda_i)^2}.$$

Which is immediately concluded if the gap between consecutive eigenvalues is decreasing.

2. Validating the theoretical results

In this section, we demonstrate the theoretical bounds derived in the main manuscript, and robustness of the proposed algorithm to outliers. In addition, we verify that time complexity of IPM is linear w.r.t. the number of original data in real simulations.

Lower Bound:
To verify the lower bound stated in Proposition 1, we ran experiment using 60,000 samples from MNIST handwritten digits dataset. Figure 1 shows the correlation between the selected sample at each selection iteration and the first right singular vector. As it can be seen, $ROM(\tilde{\boldsymbol{A}})$ is the lower bound for $max(|\tilde{\boldsymbol{A}}\boldsymbol{v}|)$. Our selection algorithm chooses the sample point corresponding to the maximum correlation of samples, with the first singular vector, i.e., max($|\tilde{\boldsymbol{A}}\boldsymbol{v}|$), and its lower bound indicated by ROM($\tilde{\boldsymbol{A}}$) in each iteration of selection. Moreover, 60,000 random samples with rank equal to 50[1] by the same dimension of MNIST are generated. The gap between the lower bound and the obtained correlation of the selected sample is huge for the random data. However, for the real structured samples of MNIST dataset a tighter lower bound is obtained.

Effect of Outliers:
The influence of outliers on the eigenvectors of the auto-correlation matrix is studied in Figure 2. The first, second, and third eigenvectors using 5000 samples of MNIST handwritten digit dataset. These eigenvectors are computed again after adding outlier from MNIST fashion dataset contaminated with noise with PSNR of +15 dB. The correlation between before and after adding outliers are calculated and plotted in this figure. As it can be seen the first eigencevtor is still 94% correlated after adding 2000 outliers, i.e., 40% of inlier samples.

The sensitivity coefficient for the first 25 eigenvector of MNIST handwritten dataset is shown in Figure 3. It is also plotted for random rank-50 data in 784 dimensional space for 60,000 synthesized samples. A small perturbation on matrix $\boldsymbol{C}$ is added and the ratio of ℓ_2 norm of changes of $\boldsymbol{v}_i$ and Frobenius norm of changes of $\boldsymbol{C}$ is plotted. As Lemma 2 suggests, s_i is an upper bound for sensitivity. Moreover, our simulations and theoretical results consistently indicate that the first eigenvector is absolutely the most robust direction.

In this experiment, we analyze the correlation of the selected submatrices before and after adding outliers for UCF 101 dataset. Projection error of data selection is reported in Table 1. The selection is performed to select 5 samples out of 2000 samples. The introduced cost function in Equation (2) of the main manuscript is evaluated for the proposed selection algorithm and three other methods and it is normalized to Frobenius norm of the matrix of original data.

[1] The rank of the matrix of all MNIST samples is approximately 50.

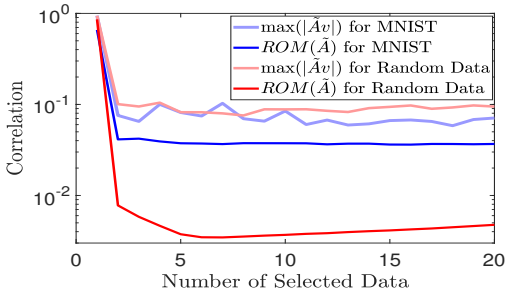


Figure 1: The maximum correlation of the first singular vector with sample points and its lower bound for MNIST handwritten dataset and a random $784 \times 60,000$ matrix with rank-50.

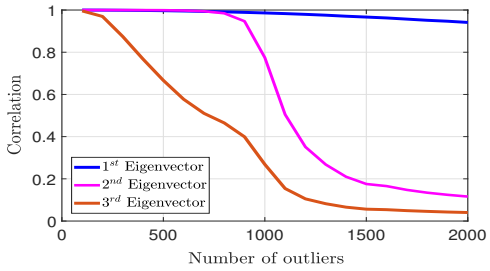


Figure 2: Robustness of the first Eigenvector to outliers.

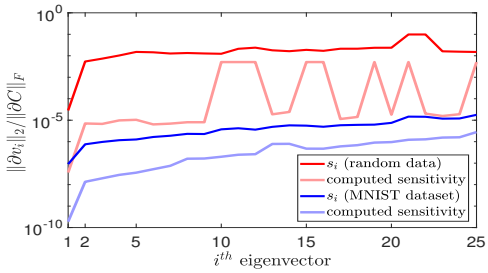


Figure 3: Sensitivity of eigenvectors w.r.t. changes in the auto-correlation matrix.

Extracted features from the pre-trained Kinetics-400 are exploited for selection where each feature is 512 dimensional. Random data from a 20 dimensional subspace are concatenated with the features as structured outliers. Our selection algorithm indicates the most accurate selection for both clean and contaminated data. After adding outliers the selection will be a different submatrix reduced from the original data. The correlation of the reduced submatrix before and after adding outliers are plotted in Figure 4. Correlation of two matrices with the same size is computed by,

$$\text{corr}(\boldsymbol{A}_1, \boldsymbol{A}_2) = \text{trace}(\boldsymbol{A}_1^T \boldsymbol{A}_2)/(\|\boldsymbol{A}_1\|_F \|\boldsymbol{A}_2\|_F).$$

DS3 shows the most consistent selection after adding outlier and IPM performs closely. However, IPM selects more accurate subset in terms of projection error as Table 1 demonstrates even after adding outliers. Please note that IPM is much faster than other algorithms. DS3 and SMRS can not be performed for large number of samples.

Finally, Figure 5 shows the running time for selecting 5 representatives from M samples. This experiment is per-

Table 1: Effect of outliers on the normalized projection error for selection of 5 representatives from extracted features of 2000 samples of UCF-101 dataset.

# Outliers	0 (clean)	100	200	300	400	500
K-mdoids	0.571	0.576	0.583	0.59	0.598	0.606
SMRS	0.638	0.643	0.649	0.655	0.661	0.669
DS3	0.603	0.608	0.614	0.621	0.628	0.639
IPM	0.557	0.569	0.582	0.598	0.615	0.627

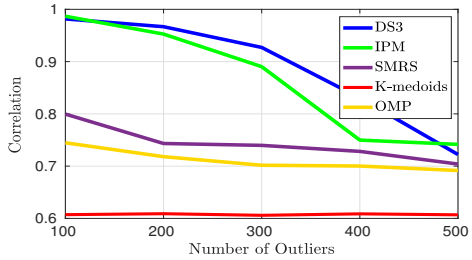


Figure 4: Correlation of the selected submatrices before and after adding outliers.

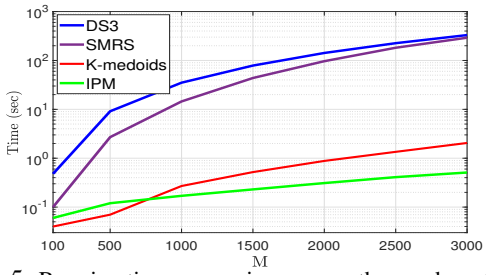


Figure 5: Running time comparison versus the number of sample points in the dataset.

formed using Intel Xeon 3.7 GHz in Matlab 2018a. This figure confirms that for large M the complexity of our algorithm is linear while DS3 and SMRS are with complexity of $O(M^3)$. K-medoids performs selection with complexity order of $O(KN(M-K)^2)$ [8].

3. Further Experiments

In this section, more experiments are provided to further investigate the performance of the proposed approach and to support the arguments presented in the main manuscript. The implementation details are the same as in the main manuscript, unless otherwise noted.

3.1. Finding Representatives for Multi-PIE Dataset

Figure 7 compares average projection error of different selection algorithms for supervised selection of K samples from each subject. Data are embedded into 200 dimensional space using PCA before performing selection. The projection error is averaged over all 249 subjects. The average error is divided by the average error of random selection in order to emphasis the gain of each selection algorithm. As it can be seen, our proposed algorithm achieves the lowest normalized projection error. In other words, the selected samples cover the space of each subject more accurately. The running time of selection from Multi-PIE dataset is shown in Figure 8. IPM performs up to 4 order of magnitude faster than DS3 algorithm while its performance is higher.

Next, an unsupervised experiment on Multi-PIE Face Database is performed, in order to show superiority of our proposed selection from a multi-class dataset, when the labels are not given. 2600 data samples from the first 5 subjects of Multi-PIE Face Database are put in a pool to select 20 samples. Figure 9 shows the number of selected samples from each subject. IPM demonstrates the most uniform selection from different subjects. Next, we consider a pool of samples consisting of 2500 samples from 25 subject, i.e., 100 samples from each subject. Figure 10 shows standard deviation of number of selected samples from each subject. IPM is the least biased algorithm toward a specific subject and its selection is more uniform.

Finally, Figure 6 shows the generated images of two more subjects in the testing set, using the trained network on the reduced dataset, as well as using the complete dataset. The implementation details are provided in Section 4.2.2 of the main manuscript. The network trained on samples selected by IPM (fourth row) is able to generate more realistic images, with fewer artifacts, compared to other selection methods (rows 1-3). Furthermore, compared to the results using all the data (row 5), it is clear that IPM-reduced dataset generates the closest results to the complete dataset.

3.2. Finding Representatives for ImageNet Dataset

Figure 11 shows the selected samples using IPM and K-medoids from different classes of the ImageNet training set. DS3 and SMRS are too computationally expensive and do not generate results for ImageNet in a tractable time. In this experiment, 5 images are selected as the representatives from each class. The implementation details are the same as given in Section 4.2.4 in the main manuscript. For each class, top row shows the images selected by IPM and bottom row shows the images selected by K-medoids. IPM-selected images are sorted by the order of selection, left-most sample being the first selected sample. Images selected by IPM are less cluttered with other objects and more representative of their corresponding classes. This leads to better classification accuracy, when the IPM-reduced representatives are used as the only labeled data available. This is demonstrated and discussed in Table 4 of the main manuscript. On the other hand, K-medoids, and other diversity-based selection methods, may select outliers or samples that may not be useful for classificaton task.

3.3. Finding Representatives for UCF-101 Dataset

Table 2 shows the classification accuracy of ResNet18 trained using the representatives selected by different meth-

Figure 6: Multi-view face generation results for two sample subjects in Multi-PIE [3] testing set using CR-GAN [7]. The network is trained on reduced training set (9 images per subject) using random selection (first row), K-medoids (second row), DS3 [2] (third row), and IPM (fourth row). The fifth row shows the results generated by the network trained on all the data (360 images per subject).

Samples per class	1	2	3	4	5	6	7	8	9	10
Random	54.6	64.7	69.2	70.5	72.9	74.0	76.0	75.6	76.0	77.0
K-medoids	61.0	67.7	69.4	70.9	71.7	72.0	72.5	75.2	73.6	73.5
DS3[2]	60.8	69.1	74.0	75.2	74.9	75.3	75.8	77.0	77.6	76.6
IPM	**65.3**	**72.6**	**74.9**	**77.6**	**77.0**	**78.5**	**78.4**	**78.4**	**79.0**	**78.2**

Table 2: Accuracy (%) of ResNet18 on UCF-101 dataset, trained using only the representatives selected by different methods. The accuracy using the full training set (9537 samples) is 82.23%.

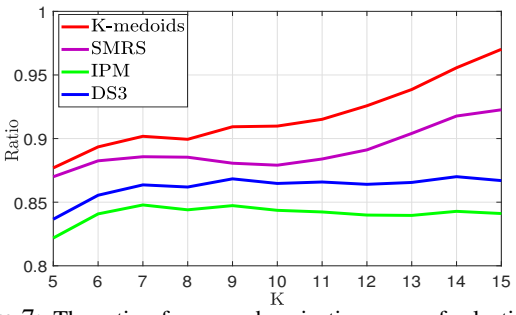


Figure 7: The ratio of averaged projection error of selection using a selection algorithm to averaged projection error of random selection for selecting K representative for each subject of Multi-PIE Face Database. Ratios are averaged for all 249 subjects of the database.

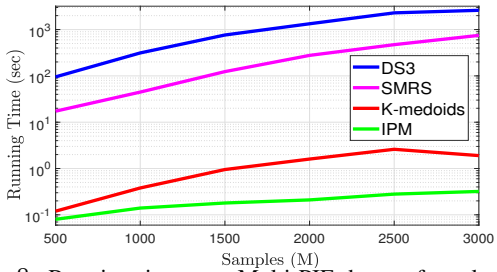


Figure 8: Running time over Multi-PIE dataset for selecting 10 representatives from a pool of M samples.

ods (extended version of Table 3 in the main manuscript). We compare IPM with DS3[2], K-medoids, and random selection as the baseline. To achieve accuracy of 77%, the closest competitor, i.e. DS3, requires 8 samples per class,

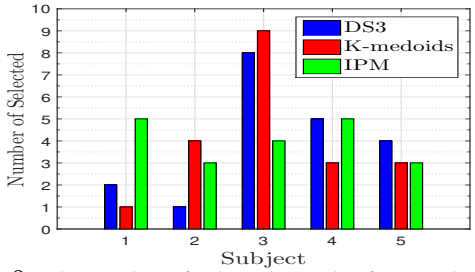


Figure 9: The number of selected samples from each subject in unsupervised selection.

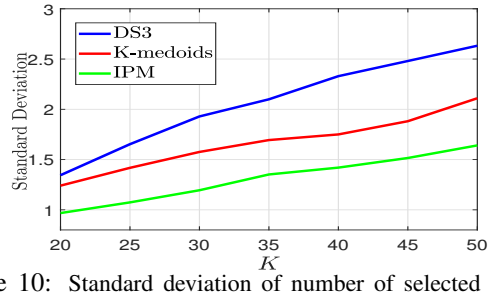
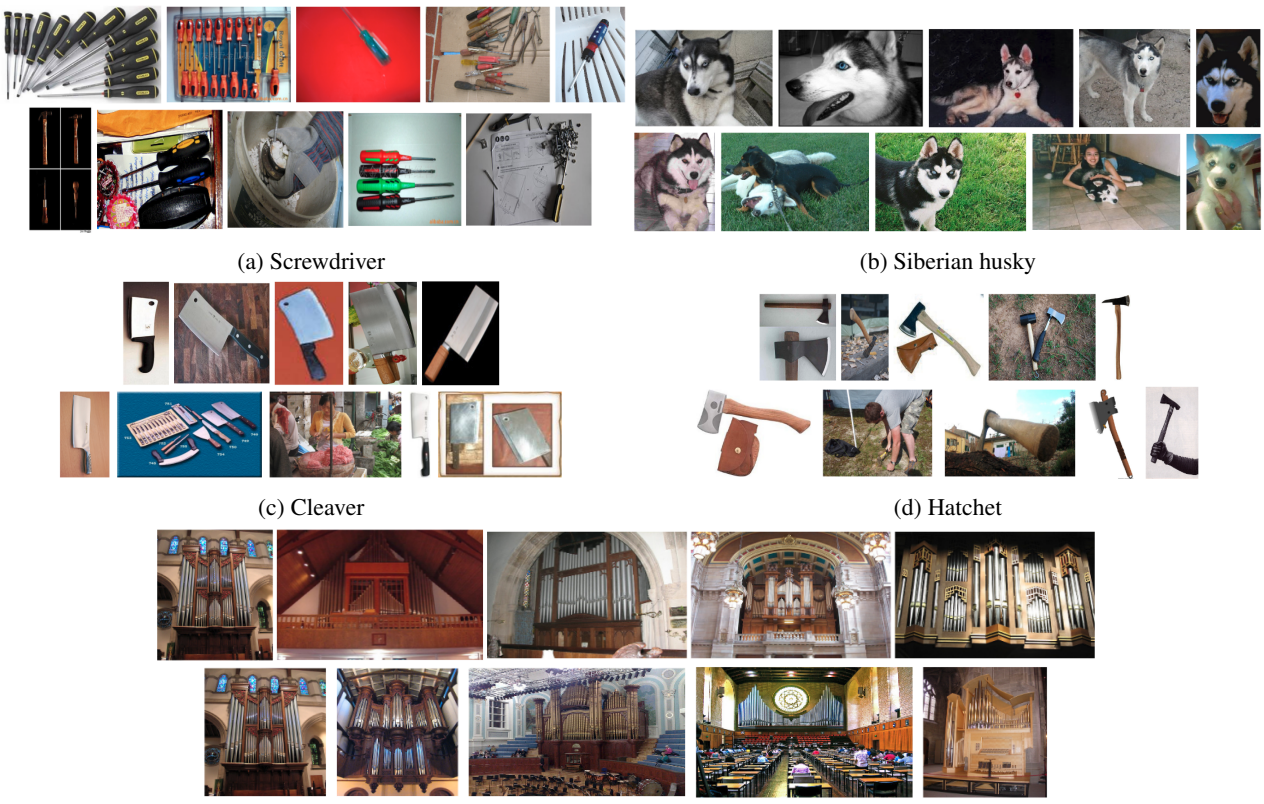


Figure 10: Standard deviation of number of selected samples from each subject. In the case of exact equal number from each subject, the standard deviation is zero.

while IPM achieves the same accuracy using half of that data. IPM adds the samples that contain the most information about the previously unseen space. This is because it selects the samples that are maximally correlated with the null space of currently selected samples. In contrast, methods such as K-medoids, that do not consider the current selected samples fail to find the most critical samples, as we collect more samples.



(a) Screwdriver

(b) Siberian husky

(c) Cleaver

(d) Hatchet

(e) Organ, pipe organ

Figure 11: Selected images by IPM (first row) and K-medoids (second row) from three sample classes of ImageNet [1]. Note that the IPM-selected samples are less cluttered with other objects, making them better representatives of the class.

This can be illustrated by t-SNE [4] visualization of the selection process. Figure 12(Left) shows the 2D embedding of the points and the decision function learned by an SVM of different randomly selected pairs of UCF-101 dataset. On the right, the decision function learned by the same classifier, trained only on a few representatives, is demonstrated. This experiment demonstrates the fact that the representatives selected by IPM contain more information about the structure of the data. Compared to other selection methods and using the same number of samples, decision function learned by the classifier trained on the IPM-selected samples looks more similar to the decision function learned from all the data. This results in more accurate classification, as reported in Table 2.

For a more qualitative investigation, Figure 13 shows frames from the first selected representative by IPM (top row) and DS3 (bottom row) for a few classes of UCF-101 dataset. In this experiment, the first selected representative by K-medoids is the same as DS3 for all the classes. In general, in the clip selected by IPM, the critical features of the action, such as barbell, violin, kayak, and bow, are more visible and/or the bounding box for the action is bigger.

3.4. Video Summarization on UT Egocentric Dataset

The ROUGE score for video summarization task on UT Egocentric dataset was reported in the main manuscript. As reported Table 5 of the main manuscript, our selection algorithm shows the closest performance to the summarization provided by human among unsupervised methods. This section provides more detail of selection from the first video of UT Egocentric dataset as an example. Figure 14 shows 24 selected scenes of the first video using IPM. The selected scenes cover the story of the whole video which is about 4 hours. Figure 15 demonstrates the textual representation of a summary created by IPM versus three different human-provided reference summaries. It is worthwhile to mention that the reference summaries contain sentences that are not in the annotation. Thus, they do not contain repetitive sentences and some of their sentences will never be selected by any selection algorithm.

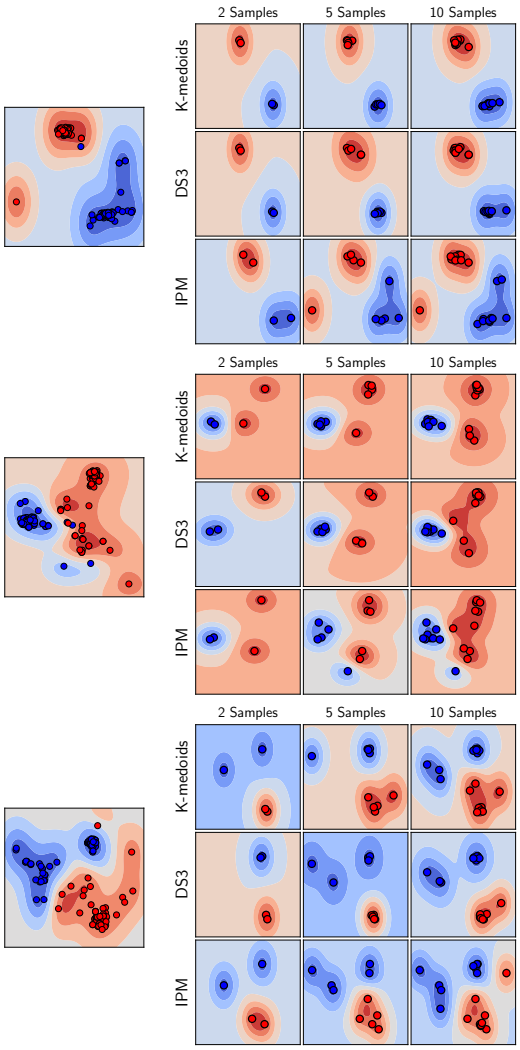
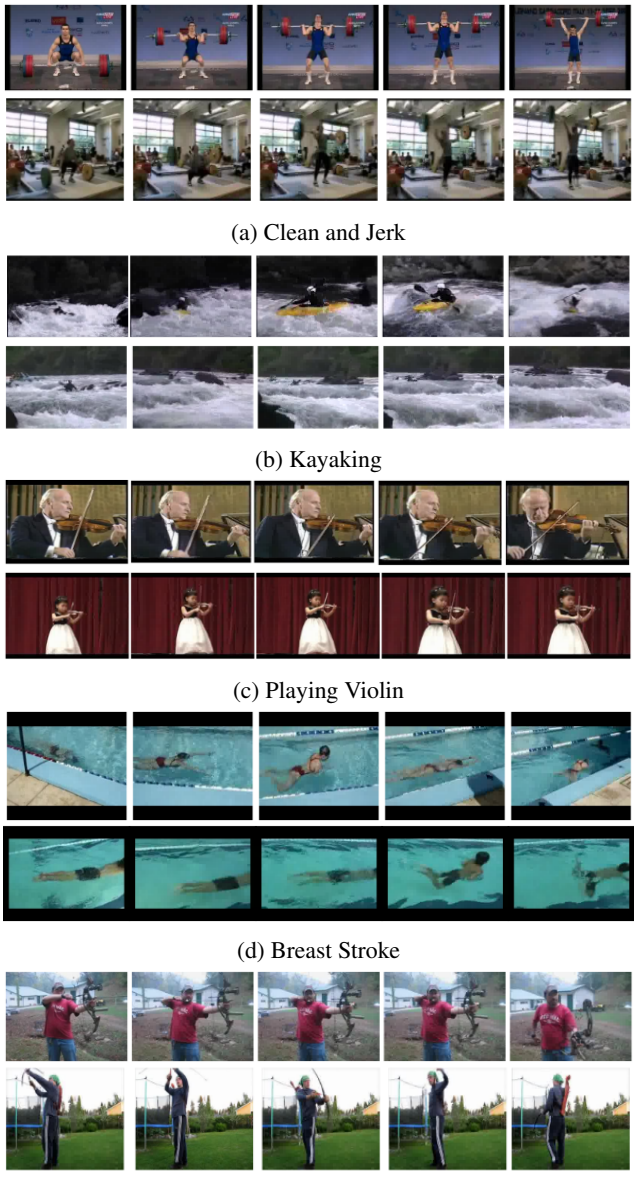


Figure 12: t-SNE visualization [4] of different randomly selected pairs of classes of UCF-101 dataset and their representatives selected by different methods. (Left) Decision function learned by using all the data. The goal of selection is to preserve the structure with only a few representatives. (Right) Decision function learned by using 2 (first column), 5 (second column), and 10 (third column) representatives per class, using K-medoids (first row), DS3 [2] (second row), and IPM (third row). IPM can capture the structure of the data better using the same number of selected samples.



(a) Clean and Jerk

(b) Kayaking

(c) Playing Violin

(d) Breast Stroke

(e) Archery

Figure 13: Frames of the selected video clips by IPM (top row) and DS3[2] (bottom row), for a few sample classes of UCF-101 dataset[6]. Different actions are more visible and/or less cluttered, in the clip selected by IPM.

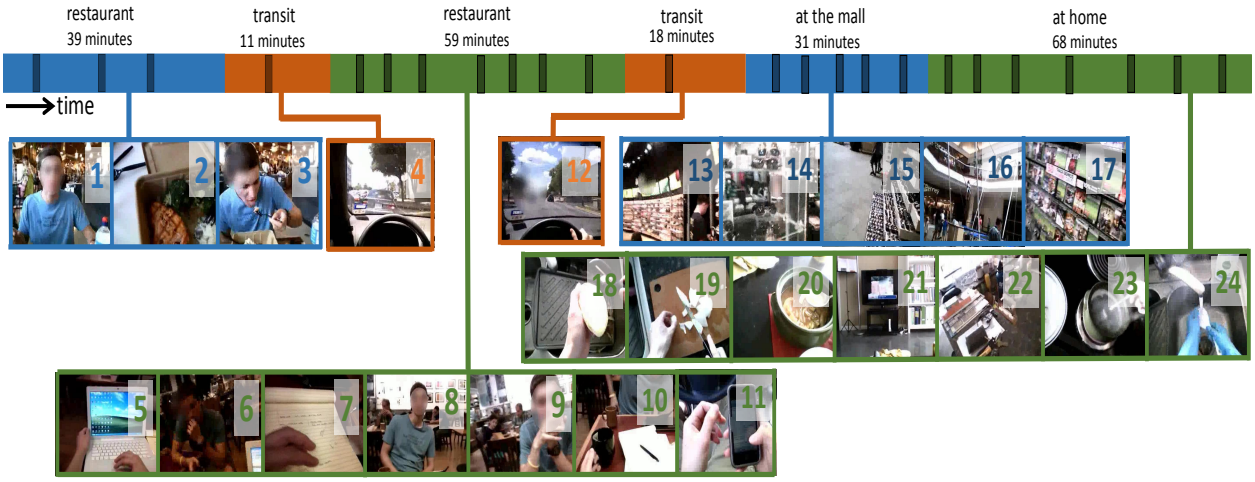


Figure 14: Two minutes summarization of the first video of UTE Egocentric dataset. The summarized annotations capture the story of the original video which is 232 minutes long. 24 clips each being 5 seconds long are selected.

	Reference 1	Reference 2	Reference 3	IPM
restaurant (39 minutes)	My friend and I sat at the table and ate a meal together. I opened the laptop. My friend and I sat at the table and talked.	I waited in line with my friend. I used the card machine to pay. I walked through the grocery store with my friend. My friend and I sat at the table and ate a meal together. My friend and I sat at the table and talked.	I waited in line with my friend. My friend and I sat at the table and ate a meal together. I walked down the street with my friend.	**I watched my friend eat. My friend and I sat at the table and ate a meal together (x2).**
transit (11 minutes)		I walked outside. I went down the escalator. I drove the car. I walked into the store. I walked through a cafe.	I walked through the store with my friend. I walked through the parking garage. I drove the car. I walked into the store.	**I drove the car.**
restaurant (59 minutes)	I used the laptop. I dipped my tea bag. I looked at the tablet. I wrote on my notepad. I drank tea.	I looked at my laptop while talking to my friend. My friend and I sat at the table and talked. I wrote on my notepad. I sat at the table with my friend and drank tea.	I put my things down on the table. I looked down at my laptop. I paid for items at the register. I sat at a table with my friend and looked at a paper. My friend and I sat at the table and talked.	**I sat in front of my laptop. I talked to my friend and looked at my laptop and phone. I wrote on my notepad. My friend and I sat at the table and talked (x3). I looked at the cell phone.**
transit (18 minutes)		I drove the car.	I drove the car. I parked the car. I walked into the mall.	**I drove the car.**
at the mall (31 minutes)	I walked into the mall. I looked at shoes on the wall. I looked at sunglasses. I watched children bounce on the trampoline. I looked at the games.	I walked into the mall. I looked at shoes on the wall. My friend and I looked at the sunglasses. I walked through the mall with my friend. I watched children bounce on the trampoline. I walked through the video game store.	I walked through the mall and talked to my friend. My friend and I walked around the mall.	**I looked at shoes on the wall. I walked in the store. I looked at the booth of glasses. I looked at the people on the bungee cords. I looked at DVDs.**
at home (68 minutes)	I used the rice cooker. I cut onions. I peeled a potato. I added a new ingredient to the cooking pot. I stirred the ingredient into the cooking pot. I sliced the cucumber. I chopped up the green onions. I added some spices to the cooking pot. I put some rice into the bowl. I added some food to my bowl. I watched television while eating my meal.	I washed dishes. I sliced onions. I peeled the potato.	I washed the dishes. I filled the pot with water from the sink and placed it on the counter. I chopped up vegetables with a knife. I stirred the ingredient into the cooking pot. I added some food to my bowl with the chopsticks. I washed the dishes in the sink.	**I sliced onions. I peeled a potato. I removed the anchovies from the pot with a fork. I ate my meal. I used the television remote. I picked up some wipes. I rinsed the debris in the sink down the drain.**

Figure 15: The details of subjective textual summarization by three Reference subjects versus the summarized annotations using IPM.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 6 2009.

[2] E. Elhamifar, G. Sapiro, and S. S. Sastry. Dissimilarity based sparse subset selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2182–2197, 2016.

[3] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, 5 2010.

[4] Laurens van der Maaten. Visualizing Data using t-SNE. *Annals of Operations Research*, 2014.

[5] C. S. RUDISILL. Derivatives of Eigenvalues and Eigenvectors for a General Matrix. *AIAA Journal*, 12(5):721–722, 5 1974.

[6] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. 12 2012.

[7] Y. Tian, X. Peng, L. Zhao, S. Zhang, and D. N. Metaxas. CR-GAN: Learning Complete Representations for Multi-view Generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 942–948, California, 7 2018. International Joint Conferences on Artificial Intelligence Organization.

[8] P. A. Vijaya, M. N. Murty, and D. K. Subramanian. Leaders–Subleaders: An efficient hierarchical clustering algorithm for large data sets. *Pattern Recognition Letters*, 25(4):505–513, 2004.