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

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

Abstract

The goal of data selection is to capture the most structural information from a set of data. This paper presents a fast and accurate data selection method, in which the selected samples are optimized to span the subspace of all data. We propose a new selection algorithm, referred to as iterative projection and matching (IPM), with linear complexity w.r.t. the number of data, and without any parameter to be tuned. In our algorithm, at each iteration, the maximum information from the structure of the data is captured by one selected sample, and the captured information is neglected in the next iterations by projection on the null-space of previously selected samples. The computational efficiency and the selection accuracy of our proposed algorithm outperform those of the conventional methods. Furthermore, the superiority of the proposed algorithm is shown on active learning for video action recognition dataset on UCF-101; learning using representatives on ImageNet; training a generative adversarial network (GAN) to generate multi-view images from a single-view input on CMU Multi-PIE dataset; and video summarization on UTE Egocentric dataset.

1. Introduction

Thanks to recent advances in computing, deep learning based systems, which employ very large numbers of inputs, have been developed in the last decade. However, processing/labeling/communication of a large number of input data has remained challenging. Therefore, novel machine learning methods that make the best use of a significantly less amount of data are of great interest. For example, active learning (AL) [26] aims at addressing this problem by training a model using a small number of labeled data, testing on the trained model, and then querying the labels of some selected data, which then are used for training a new model. In this context, preserving the underlying structure of data by a succinct format is an essential concern.

Data selection task is not trivial and possibly implies addressing an NP-hard problem (i.e., there are $\binom{M}{K}$ possibilities of choosing K distinct sample out of M available ones). This means that an optimal solution cannot be efficiently computed when the number of available data becomes excessively large. A convex relaxation of the original NP-hard problem has been suggested in terms of the D-optimal and A-optimal solutions [1, 23]. In addition to convex relaxation, a sub-modular cost function as the criterion of selection, allows us to employ much faster greedy optimization methods for selection [36]. The stochastic implementation of D-optimal solution is referred to *volume sampling* (VS), which is a fast and well-studied method. VS selects each subset of data, which are organized in the rows of a matrix, with probability proportional to the determinant (volume) of the reduced matrix. Moreover, QR decomposition with column pivoting (QRCP) and convex hull-based selection methods have been suggested for optimal data selection [10, 9]. All the mentioned methods aim to select the most diverse subset of data in an optimal sense. However, these methods do not guarantee that the un-selected samples are well-covered by the selected ones. Further, outliers are selected with a high probability using such algorithms due to their diversity, unless preprocessed by outlier detection algorithms [35]. Authors in [22] address this problem via a two-phase algorithm. There are some other efforts for outlier rejection in the selection procedure [34, 42]. However, the outlier and inlier data are not well-defined and these methods are not consistent with general data.

There is another more effective approach for subset selection, which chooses data such that the selected samples are able to approximate the rest of data accurately. This selection problem is formulated using a convex optimization problem and referred as sparse modeling representative selection (SMRS) algorithm [12]. The same goal is pursued by dissimilarity-based sparse subset selection (DS3), which is based on simultaneous sparse recovery for finding data representatives [11]. Representative approaches, such as SMRS and DS3, provide more suitable subset rather than selecting some diverse samples. However, their computa-

*indicates shared first authorship.

tional burden is not tractable for large datasets. Moreover, SMRS and DS3 algorithms utilize some parameters in their implementation, which makes their fine tuning difficult.

In order to address above issues, we propose a novel representative-based selection method, referred to as Iterative Projection and Matching (IPM). In our algorithm, at each iteration the maximum information from structure of the data is captured by one selected sample, and the captured information is neglected in the next iterations by projection on the null-space of previously selected samples. In summary, this paper makes the following contributions:

- The complexity of IPM is linear w.r.t. number of original data. Hence, IPM is tractable for larger datasets.
- IPM has no parameters for fine tuning, unlike some existing methods [11, 12]. This makes IPM dataset- and problem-independent.
- Robustness of the proposed solution is investigated theoretically.
- The superiority of the proposed algorithm is shown in different computer vision applications.

2. Problem Statement and Related Work

Let $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_M \in \mathbb{R}^N$ be M given data points of dimension N. We define an $M \times N$ matrix, $\boldsymbol{A}$, such that $\boldsymbol{a}_m^T$ is the m^{th} row of A, for $m = 1, 2, \ldots, M$. The goal is to reduce this matrix into a $K \times N$ matrix, $\boldsymbol{A}_R$, based on an optimality metric. In this section, we introduce some related work on matrix subset selection and data selection.

2.1. Selection Based on Diversity

Consider a large system of equations $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{w}$, which can be interpreted as a simple linear classifier in which $\boldsymbol{y}$ is the vector of labels, $\boldsymbol{A}$ represents the training data and $\boldsymbol{w}$ is the classifier weights. An optimal sense for data selection is to reduce this system of equations to a smaller system, $\boldsymbol{y}_R = \boldsymbol{A}_R\hat{\boldsymbol{w}}$, such that the reduced subsystem estimates the same classifier as the original system, i.e., the estimation error of $\hat{\boldsymbol{w}}$ is minimized [6].

A typical selection objective is to minimize $\mathbb{E}_\nu\{\|\boldsymbol{w} - \hat{\boldsymbol{w}}\|_2\}$, where $\mathbb{E}_\nu$ is expectation w.r.t. noise distribution of $\boldsymbol{w} - \hat{\boldsymbol{w}}$. This criterion is referred as *A-optimal* design in the literature of optimization. It is an NP hard problem, which can be solved via convex relaxation with computational complexity of $O(M^3)$ [23].

However, there are other criteria which have interesting properties. For example *D-optimal* design optimizes the determinant of a reduced matrix [23]. There are several other efforts in this area [7, 8, 27, 13, 21]. Inspired by D-optimal design, volume sampling (VS), which has received lots of attention, considers a selection probability for each subset of data, which is proportional to the determinant (volume) of the reduced matrix [27, 32, 6]. VS theory expresses that if $\mathbb{T} \subset \{1, 2, \ldots, M\}$ is any subset with cardinality K, chosen with probability proportional to det$(\boldsymbol{A}_\mathbb{T}\boldsymbol{A}_\mathbb{T}^T)$, then[1],

$$\mathbb{E}\{\|\boldsymbol{A} - \pi_\mathbb{T}(\boldsymbol{A})\|_F^2\} \leq (K+1)\|\boldsymbol{A} - \boldsymbol{A}_K\|_F^2, \tag{1}$$

where, $\pi_\mathbb{T}(\boldsymbol{A})$ is a matrix representing projection of rows of $\boldsymbol{A}$ on to the span of selected rows indexed by $\mathbb{T}$. $\mathbb{E}$ indicates expectation operator w.r.t. all the combinatorial selection of K rows of $\boldsymbol{A}$ out of M. $\boldsymbol{A}_K$ is the best rank-K approximation of $\boldsymbol{A}$ that can be obtained by singular value decomposition and $\|.\|_F^2$ is the Frobenius norm. VS is not a deterministic selection algorithm, as it gives a probability of selection for any subset of samples, and for which only a loose upper bound for the expectation of projection error is guaranteed. In contrast, in this paper a deterministic algorithm is proposed based on direct minimization of projection error using a new optimization mechanism.

2.2. Representative Selection

A method for sampling from a set of data is proposed by Elhamifar et. al. based on sparse modeling representative selection (SMRS) [12]. Their proposed cost function for data selection is the error of projecting all the data onto the subspace spanned by the selected data. Mathematically, the optimization problem in [12] can be written as,

$$\underset{|\mathbb{T}|=K}{\text{argmin}}\|\boldsymbol{A} - \pi_\mathbb{T}(\boldsymbol{A})\|_F^2. \tag{2}$$

This is an NP-hard problem. Their main contribution is solving this problem via convex relaxation. However, there is no guarantee that convex relaxation provides the best approximation for an NP-hard problem. Furthermore, such methods that try to solve the selection problem via convex programming are usually too computationally intensive for large datasets [12, 11, 31, 29]. In this paper, we propose a new fast algorithm for solving Problem (2).

Dissimilarity-based Sparse Subset Selection (DS3) algorithm selects a subset of data based on pairwise distance of all data to some target points [11]. DS3 considers a source dataset and its goal is to encode the target data according to pairwise dissimilarity between each sample of source and target datasets. This algorithm can be interpreted as the non-linear implementation of SMRS algorithm [11].

3. Iterative Projection and Matching (IPM)

In this section, an iterative and computationally efficient algorithm is proposed for approximating the solution to the NP-hard selection problem (2). The proposed algorithm iteratively finds the best direction on the unit sphere[2], and then from the available samples in dataset selects the sample with the smallest angle to the found direction.

[1] $\boldsymbol{A}_\mathbb{T}$ is the selected rows of $\boldsymbol{A}$ indexed by set $\mathbb{T}$.

[2] In unit sphere, every point corresponds to a unique direction.

Projection of all the data on to the subspace spanned by the K rows of $\boldsymbol{A}$, indexed by $\mathbb{T}$, i.e., $\pi_{\mathbb{T}}(\boldsymbol{A})$, can be expressed by a rank-K factorization, $\boldsymbol{U}\boldsymbol{V}^T$, where $\boldsymbol{U} \in \mathbb{R}^{M\times K}$, $\boldsymbol{V}^T \in \mathbb{R}^{K\times N}$, and $\boldsymbol{V}^T$ includes the K rows of $\boldsymbol{A}$, indexed by $\mathbb{T}$, and normalized to have unit length. Therefore, optimization problem (2) can be restated as

$$\underset{\boldsymbol{U},\boldsymbol{V}}{\text{argmin}}\ \|\boldsymbol{A} - \boldsymbol{U}\boldsymbol{V}^T\|_F^2 \text{ s.t. } \boldsymbol{v}_k \in \mathbb{A}, \tag{3}$$

where, $\mathbb{A} = \{\tilde{\boldsymbol{a}}_1, \tilde{\boldsymbol{a}}_2, \dots, \tilde{\boldsymbol{a}}_M\}$, $\tilde{\boldsymbol{a}}_m = \boldsymbol{a}_m/\|\boldsymbol{a}_m\|_2$, and $\boldsymbol{v}_k$ is the k^{th} column of $\boldsymbol{V}$. It should be noted that $\boldsymbol{V}^T$ is restricted to be a collection of K normalized rows of $\boldsymbol{A}$, while there is no constraint on $\boldsymbol{U}$. Assume we are to select one sample at a time, which is the best representation of all data. Since Problem (3) involves a combinatorial search and is not easy to tackle, let us modify (3) into two consecutive problems. The first sub-problem relaxes the constraint $\boldsymbol{v}_k \in \mathbb{A}$ in (3) to a moderate constraint $\|\boldsymbol{v}\| = 1$, and the second sub-problem reimposes the underlying constraint. These sub-problems are formulated as

$$(\boldsymbol{u}, \boldsymbol{v}) = \underset{\boldsymbol{u},\boldsymbol{v}}{\text{argmin}}\ \|\boldsymbol{A} - \boldsymbol{u}\boldsymbol{v}^T\|_F^2 \text{ s.t. } \|\boldsymbol{v}\| = 1, \tag{4a}$$

$$m^{(1)} = \underset{m}{\text{argmax}}\ |\boldsymbol{v}^T\tilde{\boldsymbol{a}}_m|. \tag{4b}$$

Here $m^{(1)}$ is the index of the first selected data point and $\boldsymbol{a}_{\boldsymbol{m}^{(1)}}$ is the selected sample. Subproblem (4a) is equivalent to finding the first right singular vector of $\boldsymbol{A}$. The constraint $\|\boldsymbol{v}\| = 1$ keeps $\boldsymbol{v}$ on the unit sphere to remove scale ambiguity between $\boldsymbol{u}$ and $\boldsymbol{v}$. Moreover, the unit sphere is a superset for $\mathbb{A}$ and keeps the modified problem close to the recast problem (3). After solving for $\boldsymbol{v}$ (which is not necessarily one of our data points), we find the data point that matches $\boldsymbol{v}$ the most (makes the smallest angle with $\boldsymbol{v}$) in (4b).

After selecting the first data point ($\boldsymbol{a}_{m^{(1)}}$), we project all data points onto the null space of the selected sample. This forms a new matrix $\boldsymbol{A}(\boldsymbol{I} - \tilde{\boldsymbol{a}}_{m^{(1)}}\tilde{\boldsymbol{a}}_{m^{(1)}}^T)$, where $\boldsymbol{I}$ is an identity matrix. We solve (4) with this new matrix to find the second data point. This process will continue until we select K data points. It should be noted that the null space of selected sample(s) indicates a subspace that the selected sample(s) cannot span. Therefore, the next selected data is obtained by only searching in this null space.

Algorithm 1 shows the steps of the proposed iterative projection and matching (IPM) algorithm, in which $m^{(k)}$ denotes the index of the selected data at the k^{th} iteration. IPM is a low-complexity algorithm with no parameters to be tuned. These features in addition to its superior performance (as will be shown in many scenarios in Section 4) make IPM very desirable for a wide range of applications.

Time complexity order of computing the first singular component of an $M \times N$ matrix is $O(MN)$ [4]. As the proposed algorithm only needs the first singular component for each selection, its time complexity is $O(KNM)$, which is much faster than convex relaxation-based algorithms with complexity $O(M^3)$ [23]. Moreover, IPM performs faster than K-medoids algorithm, whose complexity is of order $O(KN(M-K)^2)$ [41]. It is also worthwhile to mention that the condition that needs to be satisfied for a good performance is $K \leq N < M$. This ensures that the calculated singular vector is reliable and not impacted by noise. This condition is satisfied in subset selection scenarios, where the dataset is large, the number of selected samples is a lot less than the number of samples ($K \ll M$), and we have freedom over the dimension of the samples/features (N).

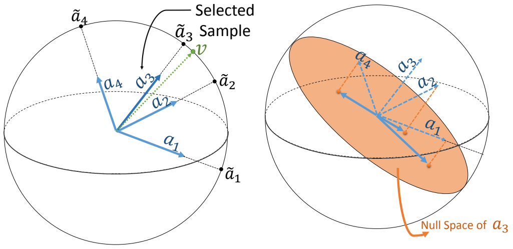


Figure 1: A toy example that illustrates the first iteration of IPM. (Left) The most matched sample with the first right singular vector, $\boldsymbol{v}$, is selected. (Right) The rest of samples are projected on the null space of the selected sample in order to continue selection in the lower dimensional subspace.

Algorithm 1 Iterative Projection and Matching Algorithm

```
Require: $\boldsymbol{A}$ and $K$
  Output: $\boldsymbol{A}_{\mathbb{T}}$
  Initialization:
  $\boldsymbol{A}^{(1)} \longleftarrow \boldsymbol{A}$
  $\mathbb{T} = \{\}$
  for $k = 1, \cdots, K$
      $v \leftarrow$ first right singular-vector of $\boldsymbol{A}^{(k)}$ by solving (4a)
      $m^{(k)} \leftarrow$ index of the most correlated data with $\boldsymbol{v}$ (4b)
      $\mathbb{T} \leftarrow \mathbb{T} \cup m^{(k)}$
      $\boldsymbol{A}^{(k+1)} \leftarrow \boldsymbol{A}^{(k)}(\boldsymbol{I} - \tilde{\boldsymbol{a}}_{m^{(k)}}\tilde{\boldsymbol{a}}_{m^{(k)}}^T)$ (null space projection)
  end
```

3.1. A Lower Bound on Maximum Correlation

In this section, we will derive a lower bound on the maximum of the absolute value of the correlation coefficient between data points $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_M$ and $\boldsymbol{v}$, when data are normalized on the unit sphere. Figure 1 shows an intuitive example for one iteration of the proposed algorithm. First, the leading singular vector is computed, and then the most correlated sample in the dataset is matched with the computed singular vector. Next, all data are projected onto the null space of the matched sample. The projected data are ready to perform one more iteration, if needed. These iterations are terminated either by reaching the desired number of selected samples or a given threshold of residual energy. Next, we present a lemma that guarantees the existence of a highly correlated sample with the first right singular vector, illustrating the fact that the selected sample will not be too bad.

Lemma 1 *Let $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_M \in \mathbb{R}^N$ be M given data points of dimension N. Let $\boldsymbol{A}$ denote an $M \times N$ matrix with $\boldsymbol{a}_m^T$ being its m^{th} row for $m = 1, 2, \ldots, M$. Let σ_1, $\boldsymbol{u}$ and $\boldsymbol{v}$ denote the first singular value, the corresponding left and right singular vectors of $\boldsymbol{A}$, respectively. Then, there exists at least one data point such that the absolute value of its inner product with $\boldsymbol{v}$ is greater than or equal to $\frac{\sigma_1}{\sqrt{M}}$. Hence, $\max_m |\boldsymbol{v}^T \boldsymbol{a}_m| \geq \frac{\sigma_1}{\sqrt{M}}$.*

The following proposition states a lower bound on the maximum of the absolute value of the correlation between data points $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_M$ and $\boldsymbol{v}$, when data are normalized on the unit sphere. First, let us define the following measure.

Definition 1 *Rank-oneness measure (ROM) of a rank R matrix $\boldsymbol{A}$ with singular values $\sigma_1, \sigma_2, \ldots, \sigma_R$ is defined as $ROM(\boldsymbol{A}) = \sqrt{\frac{\sigma_1^2}{\sum_{r=1}^R \sigma_r^2}} = \frac{\sigma_1}{\|\boldsymbol{A}\|_F}$.*

Proposition 1 *Assume the rows of $\boldsymbol{A}$ are normalized to lie on the unit sphere. There exists at least one data point, i, such that the correlation coefficient between $\boldsymbol{a_i}$ and the first right singular vector of $\boldsymbol{A}$ is greater than or equal to $ROM(\boldsymbol{A})$.*

3.2. Robustness to Perturbation

Data selection algorithms are vulnerable to outlier samples. Since outlier samples are more spread in the space of data, their span covers a wider subspace. However, the spanned subspace by outliers may not be a proper representative subspace. DS3 adds a penalty to the cost function in order to reject outliers [11]. Our proposed algorithm computes the first singular vector as the leading direction in each iteration. We show here that this direction is the most robust spectral component against changes in the data. First consider the autocorrelation matrix of data defined as, $\boldsymbol{C} = \sum_{m=1}^M \boldsymbol{a}_m \boldsymbol{a}_m^T$.

Eigenvectors of this matrix are equal to right singular vectors of $\boldsymbol{A}$. Adding a new row in $\boldsymbol{A}$ does not change the size of matrix $\boldsymbol{C}$, but perturbs this matrix. The following lemma shows the robustness of eigenvectors of $\boldsymbol{C}$ against perturbations.

Lemma 2 *Assume square matrix $\boldsymbol{C}$ and its spectrum $[\lambda_i, v_i]$. Then, the following inequality holds,*

$$\|\partial v_i\|_2 \leq \sqrt{\sum_{j \neq i} \frac{1}{(\lambda_i - \lambda_j)^2}} \|\partial C\|_F.$$

Definition 2 *The sensitivity coefficient of the i^{th} eigenvector of a square matrix is defined by, $s_i \triangleq \sqrt{\sum_{j \neq i} \frac{1}{(\lambda_i - \lambda_j)^2}}$.*

It is easy to show that $s_1 < s_2$. Based on Lemma 2 and this definition the following proposition suggests a condition to satisfy $s_1 < s_i, \ \forall i \geq 2$.

Proposition 2 *Assume square matrix $\boldsymbol{C}$ and its spectrum $[\lambda_i, v_i]$, where the gap between consecutive eigenvalues is decreasing. Then, $s_1 < s_i, \ \forall i \geq 2$.*

The proofs of Propositions and Lemmas in this section are presented in the supplementary material. Moreover, the results of Proposition 1 and 2 are also verified in supplementary material.

4. Applications of IPM

To validate our theoretical investigation and to empirically demonstrate the behavior and effectiveness of the proposed selection technique, we have performed extensive sets of experiments considering several different scenarios. We divide our experiments into three different subsections. In Section 4.1, we use our algorithm in the *active learning* setting and show that IPM is able to reduce the labelling cost significantly, by selecting the most informative unlabeled samples. Next, in Section 4.2, we show the effectiveness of IPM in selecting the most informative representatives, by training the classifier using only a few representatives from each class. Lastly, in Section 4.3, the application of IPM for video summarization is exhibited. In addition, we investigate the robustness and other performance metrics, such as projection error and running time, of different selection methods and verify our theoretical results in the supplementary material[3].

4.1. Active Learning

Active learning aims at addressing the costly data labeling problem by iteratively training a model using a small number of labeled data, and then querying the labels of some selected data, using an acquisition function.

In active learning, the model is initially trained using a small set of labeled data (the initial training set). Then, the acquisition function selects a few points from the pool of unlabeled data, asks an oracle (often a human expert) for the labels, and adds them to the training set. Next, a new model is trained on the updated training set. By repeating these steps, we can collect the *most informative* samples, which often result in significant reductions in the labeling cost. Now, the fundamental question in active learning is: Given a fixed labeling budget, what are the best unlabeled data instances to be selected for labeling for the best performance?

[3] Code for IPM is available at `cwnlab.eecs.ucf.edu/ipm/`

Mean samples/class	2	3	4	5	6	7	8
Random	60.1 ± 0.7	65.1 ± 1.2	68.2 ± 1.7	69.9 ± 1.4	71.7 ± 0.6	73.0 ± 0.6	74.8 ± 0.5
Spectral Clustering	62.3 ± 1.9	66.9 ± 1.1	68.1 ± 0.7	68.9 ± 0.3	70.8 ± 0.9	71.0 ± 2.2	71.6 ± 0.1
K-medoids	60.1 ± 2.2	65.3 ± 1.0	68.4 ± 1.6	69.2 ± 0.5	72.3 ± 0.7	73.6 ± 0.4	74.5 ± 0.6
OMP	64.2 ± 0.6	66.6 ± 0.7	70.8 ± 1.5	71.7 ± 0.4	74.3 ± 0.7	74.3 ± 0.3	75.4 ± 0.2
DS3 [11]	64.0 ± 1.5	66.5 ± 0.7	67.8 ± 1.2	68.3 ± 0.5	69.6 ± 1.1	70.9 ± 1.3	71.9 ± 0.9
Uncertainty [15]	59.5 ± 0.4	66.7 ± 1.6	69.4 ± 1.7	71.5 ± 1.5	73.9 ± 0.3	75.5 ± 0.7	75.9 ± 1.1
IPM	$\mathbf{64.6 \pm 0.7}$	68.7 ± 0.3	72.2 ± 1.0	73.4 ± 0.9	74.3 ± 0.4	74.7 ± 1.4	75.3 ± 0.6
IPM + Uncertainty	64.3 ± 0.4	$\mathbf{69.4 \pm 0.8}$	$\mathbf{72.8 \pm 1.0}$	$\mathbf{73.8 \pm 0.9}$	$\mathbf{76.2 \pm 1.0}$	$\mathbf{76.3 \pm 0.3}$	$\mathbf{77.9 \pm 0.2}$

Table 1: Classification accuracy (%) for action recognition on UCF-101, at different active learning cycles. The initial training set (cycle 1) is the same for all the methods. The accuracy for cycle 1 is 54.2% and the accuracy using the full training set (9537 samples) is 82.23%.

In many active learning frameworks, new data points are selected based on the model uncertainty. However, the effect of such selection only kicks in after the size of the training set is large enough, so we can have a reliable uncertainty measure. In this section, we show that the proposed selection method can effectively find the best representatives of the data and outperforms several recent uncertainty-based and algebraic selection methods.

In particular, we study IPM for active learning of video action recognition, using the 3D ResNet18 architecture[4], as described in [20]. The experiments are run on UCF-101 human action dataset [37], and the network is pretrained on Kinetics-400 dataset [24]. We provide the results on split 1.

To ensure that at least one sample per class exists in the training set, for the initial training, one sample per class is selected randomly and the fully-connected layer of the classifier is fine tuned. Then, at each active learning cycle, one sample per class is selected, without the knowledge of the labels, and added to the training set. Next, using the updated training set, the fully connected layer of the network is fine tuned for 60 epochs, using learning rate of 10^{-1}, weight decay of 10^{-3}, and batch size of 24 on 2 GPUs. Rest of the implementation and training settings are the same as [20]. Note that, in this experiment, fine-tuning is only performed to train the fully connected layer, because it achieved the best accuracy during the preliminary investigation for very small training sets, which is the scope of this experiment.

The selection is performed on the convolutional features extracted from the last convolutional layer of the network. Table 1 shows the accuracy of the trained network at each active learning cycle for different selection methods. The high computational complexity of DS3 prevents its implementation on all the data[5] [11]. So, we provide the results for DS3 only for the clustered version, meaning that one sample per cluster is selected using DS3 (clusters are obtained using spectral clustering). For spectral clustering results, the extracted features are clustered into 101 clusters, and one sample from each cluster is selected randomly. Furthermore, OMP, which stands for Orthogonal Matching Pursuit, selects the samples that are most correlated with the null space of the selected samples [40, 2]. The OMP approach is very sensitive to the outliers. Random outliers have low correlation with the samples and therefore a high correlation with the null space of the selected samples.

For uncertainty-based selection, Bayesian active learning [15, 14] is utilized. For that, a dropout unit with parameter 0.2 is added before the fully-connected layer and the uncertainty measure is computed by using 10 forward iterations (following the implementation in [14]). In our experiments, we use variation ratio[6] as the uncertainty metric, which is shown to be the most reliable metric among several well-known metrics [15]. Also, for a fair comparison, the initial training set is the same for all the experiments at each run.

It is evident that, during the first few cycles, since the classifier is not able to generate reliable uncertainty score, uncertainty-based selection does not lead to a performance gain. In fact, random selection outperforms uncertainty-based selection. On the other hand, IPM is able to select the critical samples. In the first few active learning cycles, IPM is constantly outperforming other methods, which translates into significant reductions in labeling cost for applications such as video action recognition.

As the classifier is trained with more data, it is able to provide us with better uncertainty scores. Thus to enjoy the benefits of both IPM and uncertainty-based selection, we can use a compound selection criterion. For the extremely small datasets, samples should be selected only using IPM. However, as we collect more data, the uncertainty score should be integrated into the decision making process. Our proposed selection algorithm, unlike other methods, easily lends itself to such modification. At each selection iteration, instead of selecting the most correlated data with $\boldsymbol{v}$ (line 3 in Algorithm 1), we can select the samples based on the following criterion:

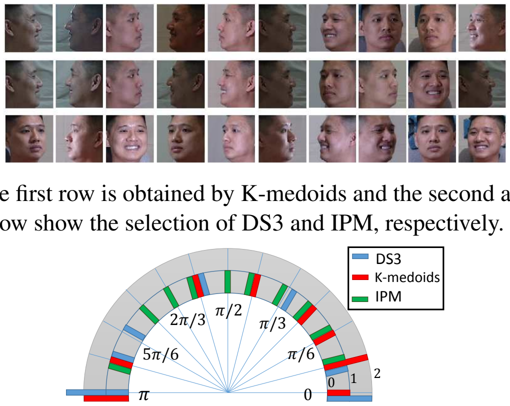
$$m^* = \arg\max_{m} \alpha \, |\boldsymbol{v}^T \tilde{\boldsymbol{a}}_m| + (1-\alpha) \, q(\boldsymbol{a}_m),$$

where $q(.)$ is an uncertainty measure, e.g. variation ratios. Parameter α determines the relative importance of the IPM

[4] We use the code provided by the authors at https://github.com/kenshohara/3D-ResNets-PyTorch

[5] We use the code provided by the authors at http://www.ccs.neu.edu/home/eelhami/codes.htm

[6] Variation ratio of x is defined as $1 - \max_y p(y|x)$. which measures lack of confidence.



(a) The first row is obtained by K-medoids and the second and the third row show the selection of DS3 and IPM, respectively.

(b) Angles of the selected images. K-medoids selects 8 different angles. DS3 algorithm selects from 7 angles and our proposed IPM selects the maximum possible 10 distinguished angles.

Figure 2: Selection of 10 representatives out of 520 images of a subject and their corresponding angles.

metric versus the uncertainty metric. To gradually increase the impact of $q(.)$, as the model becomes more reliable, we start by setting $\alpha = 1$ and multiply it by decay rate of 0.95 at each active learning cycle. This compound selection criteria leads to better results for larger dataset sizes.

4.2. Learning Using Representatives

In this experiment, we consider the problem of learning using representatives. We find the best representatives for each class and use this reduced training set for learning. Finding representatives reduces the computation and storage requirements, and can even be used for tasks such as clustering. In the ideal case, if we collect the samples that contain enough information about the distribution of the whole dataset, the learning performance would be very close to the performance using all the data.

4.2.1 Finding Representatives for Multi-PIE Dataset

Here, we present our experimental results on CMU Multi-PIE Face Database [17]. We use 249 subjects from the first session with 13 poses, 20 illuminations, and two expressions. Thus, there are $13 \times 20 \times 2$ images per subject. Figure 2a shows 10 selected images from 520 images of a subject. As it can be seen, the results of K-medoids and DS3 algorithms are concentrated on side views, while our selection provides images from more diverse angles. Figure2b highlights this by showing the angles of selected images of each algorithm. IPM selects from 10 different angles, while the selected images by DS3 and K-medoids contain repetitious angles. Figure 3 shows the performance of different selection algorithms in terms of normalized projection error and running time. It is evident that our proposed approach finds a better minimizer for Problem defined in equation (2) and is able to do so in orders of magnitude less time.

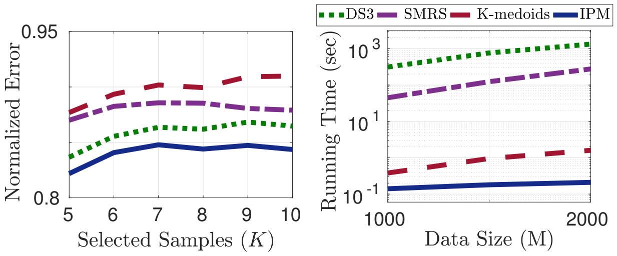


Figure 3: Performance of different methods for minimizing the cost function of representative selection in equation (2). (Left) The ratio of projection error using selection algorithms to projection error of random selection for selecting K representatives from each subject, averaged over all the subjects. (Right) Running time of different algorithms versus number of input samples.

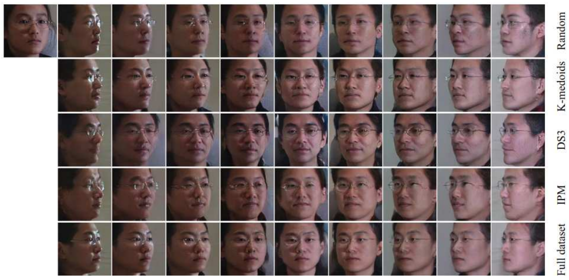


Figure 4: Multi-view face generation results for a sample subject in testing set using CR-GAN [39]. The network is trained on reduced training set (9 images per subject) using random selection (first row), K-medoids (second row), DS3 [11] (third row), and IPM (fourth row). The fifth row shows the results generated by the network trained on all the data (360 images per subject). IPM-reduced dataset generates closest results to the complete dataset.

4.2.2 Representatives To Generate Multi-view Images Using GAN

Next, to investigate the effectiveness of the proposed selection, we use the selected samples to train a generative adversarial network (GAN) to generate multi-view images from a single-view input. For that, the GAN architecture proposed in [39] is employed. Following the experiment setup in [39], only 9 poses between $\frac{\pi}{6}$ and $\frac{5\pi}{6}$ are considered. Furthermore, the first 200 subjects are for training and the rest are for testing. Thus, the total size of the training set is $72,000$, 360 per subject. All the implementation details are same as [39], unless otherwise is stated[7].

We select only 9 images from each subject (1800 total), and train the network with the reduced dataset for 300 epochs using the batch size of 36. Figure 4 shows the generated images of a subject in the testing set, using the trained network on the reduced dataset, as well as using the complete dataset. 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

[7] We use the code provided by the authors at `https://github.com/bluer555/CR-GAN`

Method	Random	K-Medoids	DS3	IPM
9 images / subject	0.5616	0.5993	0.6022	**0.553**
360 images / subject	0.5364			

Table 2: Identity dissimilarities between real and generated images by network trained on reduced (using different selection methods) and complete dataset.

Samples / Class	1	2	3	4	5	6
Random	54.6	64.7	69.2	70.5	72.9	74.0
K-medoids	61.0	67.7	69.4	70.9	71.7	72.0
OMP	51.1	64.6	70.7	72.8	73.0	74.5
DS3[11]	60.8	69.1	74.0	75.2	74.8	75.3
IPM	**65.3**	**72.6**	**74.9**	**77.6**	**77.0**	**78.5**

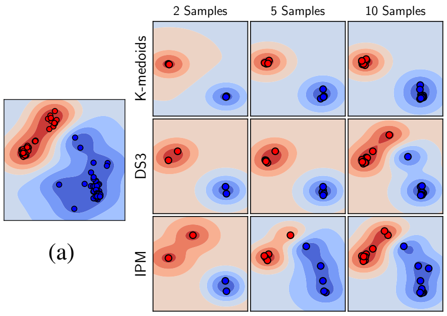
Table 3: Accuracy (%) of ResNet18 on UCF-101 dataset, trained using only the representatives selected by different methods. The accuracy using the full training set (9537 samples) is 82.23%.

(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. This is because, as demonstrated in Figure 2, samples selected by IPM cover more angles of the subject, leading better training of the GAN. See supplementary material for further experiments and sample outputs.

For a quantitative performance investigation, we evaluate the identity similarities between the real and generated images. For that, we feed each pair of real and generated images to a ResNet18[8], trained on MS-Celeb-1M dataset [18], and obtain 256-dimensional features. ℓ_2 distances of features correspond to the face dissimilarity. Table 2 shows the normalized ℓ_2 distances between the real and generated images, averaged over all the images in the testing set. Our method outperforms other selection methods in this metric as well. Thus, from Figure 4 (qualitative) and Table 2 (quantitative), we can conclude that the IPM-reduced training set contains more information about the complete set, compared to other selection methods.

4.2.3 Finding Representatives for UCF-101 Dataset

Here, similar to Section 4.1, we use a 3D ResNet18 classifier pretrained on Kinetics-400 dataset, and the selection algorithms are performed on feature space generated by the output of the last convolutional layer. To find the representatives, we use the selection methods to sequentially find the most informative representatives from each class. After selecting the representatives, the fully connected layer of the network is finetuned in the same manner as described in Section 4.1. Table 3 shows the performance of different selection methods for different numbers of representatives per class. As more samples are collected, the performance gap among different methods, including random, decreases. This is expected, since finding only one representative for each class is a much more difficult task, compared to choosing many, e.g. 6, representatives.



(a)

(b)

Figure 5: t-SNE visualization [25] of two randomly selected classes of UCF-101 dataset and their representatives selected by different methods. (a) Decision function learned by using all the data. The goal of selection is to preserve the structure with only a few representatives. (b) Decision function learned by using 2 (first column), 5 (second column), and 10 (third column) representatives per class, using K-medoids (first row), DS3 [11] (second row), and IPM (third row). IPM can capture the structure of the data better using the same number of selected samples.

Images per Class	1 (0.08%)	5 (0.4%)	10 (0.8%)	50 (4%)
Random	3.18	8.71	12.97	25.61
K-Medoids	11.78	17.01	17.56	26.86
IPM	12.50	21.69	25.26	30.77

Table 4: Top-1 classification accuracy (%) on ImageNet, using selected representatives from each class. Accuracy using all the labeled data (1.2M samples) is 46.86%. Numbers in () show the size of the selected representatives as a % of the full training set.

Using only one representative selected by IPM, we can achieve a classification accuracy of 65.3%, which is more than 10% improvement compared to random selection and more than 4% improvement compared to other competitors.

Figure 5 shows the t-SNE visualization [25] of the selection process for two randomly selected classes of UCF-101. To visualize the structure of the data, the contours represent the decision function of an SVM trained in this 2D space. Selection is performed on the original 512-dimensional feature space. This experiment illustrates that each IPM sample contains new structural information, as the selected samples are far away from each other in the t-SNE space, compared to other methods. Moreover, it is evident that as we collect more samples, the structure of the data is better captured by the samples selected by IPM, compared to other methods selecting the same number of representatives. The decision boundaries of the classifier trained on 5 IPM-selected samples look very similar to the boundaries learned from all the data. This leads to significant accuracy improvements, as already discussed and exhibited in Table 3.

4.2.4 Finding Representatives for ImageNet

In this section, we use ImageNet dataset [5] to show the effectiveness of IPM in selecting the representatives for im-

[8] We use the naive ResNet18 architecture as described in [3].

Method	F-measure	Recall
Selection Methods (Unsupervised)		
Random	26.30	23.73
Uniform	28.68	25.76
K-medoids	30.11	27.30
DS3	30.13	27.34
IPM	**31.53**	**29.09**
Supervised Summarization Methods		
SeqDPP [16]	28.87	26.83
Submod-V [19]	29.35	27.43
Submod-V+ [33]	34.15	31.59

Table 5: F-measure and recall scores using ROUGE-SU metric for UT Egocentric video summarization task. Results are reported for several supervised and unsupervised methods.

age classification task. For that, first, we extract features from images in an unsupervised manner, using the method proposed in [43]. We then perform selection in the learned 128-dimensional space and perform k-nearest neighbors (k-NN) using the learned similarity metric, following the experiments in [43][9]. Here, we show that we can learn the feature space and the similarity metric in an unsupervised manner, as there is no shortage of unlabeled data, and use only a few labeled representatives to classify the data.

Due to the volume of this dataset, selection methods based on convex-relaxation, such as DS3 [11] and SMRS [12], fail to select class representatives in a tractable time (as discussed before and shown in Figure 3 for Multi-PIE dataset). Table 4 shows the top-1 classification accuracy for the testing set using k-NN. Using less than 1% of the labels, we can achieve an accuracy of more than 25%, showing the potential benefits of the proposed approach for dataset reduction. Classification accuracy of k-NN, using the learned similarity metric, reflects the representativeness of the selected samples, thus highlighting the fact that IPM-selected samples preserve the structure of the data fairly well.

4.3. Video Summarization

In this section, we evaluate the performance of the proposed selection algorithm on the video summarization task. The goal is to select key frames/clips and create a video summary, such that it contains the most essential contents of the video. We evaluate our approach on UT Egocentric (UTE) dataset [45, 30]. It contains 4 first-person videos of 3-5 hours of daily activities, recorded in an uncontrolled environment. Authors in [44] have provided text annotations for each 5-second segment of the video, as well as human-provided reference text summaries for each video. Following [33, 19, 44], the performance is evaluated in text domain. For that, a text summary is created by concatenating the text annotations associated with the selected clips. The generated summaries are compared with the reference summaries using the ROUGE metric [28]. As in prior work, we report f-measure and recall using the ROUGE-SU score with the same parameters as in [33, 19, 44].

Table 5 provides the results for two-minute-long summaries (24 5-second samples), generated by different methods. To generate results using K-medoids, DS3, and IPM, we use 1024-dimensional feature vectors extracted using GoogleNet [38], as described in [46]. Then, the features are clustered into 24 clusters using K-means and one sample is selected from each cluster using different selection techniques. The results are the mean results over all the 4 videos and over 100 runs. Furthermore, for the supervised methods, the results are as reported in [33]. *The proposed unsupervised selection method, IPM, is the closest competitor to the state-of-art supervised method proposed in [33], outperforming other unsupervised methods and some of the supervised methods*. These supervised methods split the dataset into training, and testing sets and use reference text or video summaries of the training set to learn to summarize the videos from the test set. This experiment demonstrates the strength of IPM and the potential benefits of employing it in more advanced unsupervised or supervised schemes.

5. Conclusions

A novel data selection algorithm, referred to as Iterative Projection and Matching (IPM) is presented, that selects the most informative data points in an iterative and greedy manner. Interestingly, we show that our greedy approach, with linear complexity wrt the dataset size, is able to outperform state-of-the-art methods, which are based on convex relaxation, in several performance metrics such as projection error and running time. Furthermore, the effectiveness and compatibility of our approach are demonstrated in a wide array of applications such as active learning, video summarization, and learning from representatives. This motivates us to further investigate the potential benefits and applications of IPM in other computer vision problems.

6. Acknowledgements

This research is based upon work supported in parts by the National Science Foundation under Grants No. 1741431 and CCF-1718195 and the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. D17PC00345. The views, findings, opinions, and conclusions or recommendations contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the NSF, ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

[9] We use the feature space generated by the ResNet50 backbone, as provided in `https://github.com/zhirongw/lemniscate.pytorch`

References

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[2] T. T. Cai and L. Wang. Orthogonal Matching Pursuit for Sparse Signal Recovery With Noise. *Information Theory, IEEE Transactions on*, 57(7):4680–4688, 7 2011.

[3] K. Cao, Y. Rong, C. Li, X. Tang, and C. Change Loy. Pose-Robust Face Recognition via Deep Residual Equivariant Mapping. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] P. Comon and G. H. Golub. Tracking a few extreme singular values and vectors in signal processing. *Proceedings of the IEEE*, 78(8):1327–1343, 1990.

[5] 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.

[6] M. Derezinski and M. Warmuth. Subsampling for Ridge Regression via Regularized Volume Sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 716–725, 2018.

[7] A. Deshpande and L. Rademacher. Efficient volume sampling for row/column subset selection. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 329–338. IEEE, 2010.

[8] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1117–1126. Society for Industrial and Applied Mathematics, 2006.

[9] S. Ding, X. Nie, H. Qiao, and B. Zhang. A fast algorithm of convex hull vertices selection for online classification. *IEEE transactions on neural networks and learning systems*, 29(4):792–806, 2018.

[10] J. A. Duersch and M. Gu. Randomized QR with column pivoting. *SIAM Journal on Scientific Computing*, 39(4):C263–C291, 2017.

[11] 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.

[12] E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1600–1607. IEEE, 2012.

[13] A. K. Farahat, A. Elgohary, A. Ghodsi, and M. S. Kamel. Greedy column subset selection for large-scale data sets. *Knowledge and Information Systems*, 45(1):1–34, 2015.

[14] Y. Gal and Z. Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *PMLR*, 2016.

[15] Y. Gal, R. Islam, and Z. Ghahramani. Deep Bayesian Active Learning with Image Data. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192, International Convention Centre, Sydney, Australia, 2017. PMLR.

[16] B. Gong, W.-L. Chao, K. Grauman, and F. Sha. Diverse Sequential Subset Selection for Supervised Video Summarization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2069–2077. Curran Associates, Inc., 2014.

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

[18] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: Challenge of Recognizing One Million Celebrities in the Real World. *Electronic Imaging*, 2016(11):1–6, 2 2016.

[19] M. Gygli, H. Grabner, and L. Van Gool. Video summarization by learning submodular mixtures of objectives. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3090–3098. IEEE, 6 2015.

[20] K. Hara, H. Kataoka, and Y. Satoh. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 18–22, 2017.

[21] M. Joneidi, A. Zaeemzadeh, and N. Rahnavard. Dynamic Sensor Selection for Reliable Spectrum Sensing via E-optimal Criterion. In *2017 IEEE 14th International Conference on Mobile Adhoc and Sensor Systems*, Orlando, 2017. IEEE Computer Society.

[22] M. Joneidi, A. Zaeemzadeh, B. Shahrasbi, G.-J. Qi, and N. Rahnavard. E-optimal Sensor Selection for Compressive Sensing-based Purposes. *IEEE Transactions on Big Data*, page To Appear in, 2018.

[23] S. Joshi and S. Boyd. Sensor Selection via Convex Optimization. *IEEE Transactions on Signal Processing*, 57(2):451–462, 2 2009.

[24] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The Kinetics Human Action Video Dataset. 5 2017.

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

[26] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc., 1994.

[27] C. Li, S. Jegelka, and S. Sra. Polynomial time algorithms for dual volume sampling. In *Advances in Neural Information Processing Systems*, pages 5038–5047, 2017.

[28] C. Y. Lin. Rouge: A package for automatic evaluation of summaries. *Annual Meeting of the Association for Computational Linguistics*, 2004.

[29] H. Liu, Y. Liu, and F. Sun. Robust Exemplar Extraction Using Structured Sparse Coding. *IEEE Transactions on Neural Networks and Learning Systems*, 26(8):1816–1821, 8 2015.

[30] Z. Lu and K. Grauman. Story-Driven Summarization for Egocentric Video. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2714–2721. IEEE, 6 2013.

[31] J. Meng, H. Wang, J. Yuan, and Y.-P. Tan. From Keyframes to Key Objects: Video Summarization by Representative Object Proposal Selection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1039–1048. IEEE, 6 2016.

[32] A. Nikolov, M. Singh, and U. T. Tantipongpipat. Proportional Volume Sampling and Approximation Algorithms for A-Optimal Design. *arXiv preprint arXiv:1802.08318*, 2018.

[33] B. A. Plummer, M. Brown, and S. Lazebnik. Enhancing Video Summarization via Vision-Language Embedding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1052–1060. IEEE, 7 2017.

[34] M. Rahmani and G. K. Atia. Robust and Scalable Column/Row Sampling from Corrupted Big Data. In *ICCV Workshops*, pages 1818–1826, 2017.

[35] M. Sedghi, G. Atia, and M. Georgiopoulos. Robust Manifold Learning via Conformity Pursuit. *IEEE Signal Processing Letters*, 26(3):425–429, 3 2019.

[36] M. Shamaiah, S. Banerjee, and H. Vikalo. Greedy sensor selection: Leveraging submodularity. In *Decision and Control (CDC), 2010 49th IEEE Conference on*, pages 2572–2577, 12 2010.

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

[38] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9. IEEE, 6 2015.

[39] 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.

[40] J. A. Tropp and A. C. Gilbert. Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 12 2007.

[41] 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.

[42] H. Wang, Y. Kawahara, C. Weng, and J. Yuan. Representative selection with structured sparsity. *Pattern Recognition*, 63:268–278, 2017.

[43] Z. Wu, Y. Xiong, X. Y. Stella, and D. Lin. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[44] S. Yeung, A. Fathi, and L. Fei-Fei. VideoSET: Video Summary Evaluation through Text. 6 2014.

[45] Yong Jae Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1346–1353. IEEE, 6 2012.

[46] K. Zhang, W. L. Chao, F. Sha, and K. Grauman. Video summarization with long short-term memory. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016.