SChISM: Semantic Clustering via Image Sequence Merging for Images of Human-Decomposition

Sara Mousavi Dylan Lee Tatianna Griffin Kelley Cross Dawnie Steadman
Audris Mockus
University of Tennessee, Knoxville
{mousavi, dlee97, tgriff25, kcross12}@vols.utk.edu, {dsteadma, audris}@utk.edu

Abstract

In many domains, large image collections are key ways in which information about relevant phenomena is retained and analyzed, yet it remains challenging to use such data in research and practice. Our aim is to investigate this problem in the context of a forensic unlabeled dataset of over 1M human decomposition photos. To make this collection usable by experts, various body parts first need to be identified and traced through their evolution despite their distinct appearances at different stages of decay from "fresh" to "skeletonized". We developed an unsupervised technique for clustering images that builds sequences of similar images representing the evolution of each body part through stages of decomposition. Evaluation of our method on 34,476 human decomposition images shows that our method significantly outperforms the state of the art clustering method in this application.

1. Introduction

Evolving images with conceptual likeness when the similarity is declining over time are not uncommon, yet such data confounds clustering approaches that rely on measures of image similarity as the early stages of the same conceptual object may bear no visual resemblance to the late stages. In the case of human decomposition, a hand appears very different in the fresh stage compared to when it is decayed (Figure 1). Supervised techniques might fare better in such situations, but the creation of the labels may have prohibitive costs exacerbated by the inability to do crowdsourcing when domain experts (as in our case forensic anthropologists) may be scarce.

This paper introduces a technique for clustering evolving images in the context of human decomposition data. Specifically, the goal of this work is to jointly cluster body parts and decomposition stages within subjects and to trace them through their decay process, which spans from "fresh" to "skeletal". Such an unsupervised approach, if successful, would reduce the manual labeling task required to extract domain specific features needed for key forensic tasks such as time of death estimation and, more generally, human decomposition research and analysis [22, 23].

The key point of clustering is to segment a large collection of observations into a smaller set of groups of similar observations, which can help in understanding large datasets. Traditionally, clustering methods group images based on the similarity of features extracted from them. With adequate feature representations, image clustering methods have achieved good results [7, 12, 13, 19] on popular image datasets such as ImageNet, MNIST, COIL100, and VOC2007 [8, 9, 26, 11]. Guérin et al. [13] used pre-trained CNNs on common datasets such as ImageNet to map images to feature representations and then clustered them. Other unsupervised frameworks introduce clustering losses to jointly learn ConvNet features and image clusters in an end-to-end manner [3, 10, 18, 32, 34, 4]. However, we are not aware of any work that clusters image datasets with evolving content based on their semantic similarity in an unsupervised fashion.

In human decomposition data, although images representing same objects may change only slightly from one time to the next, these small changes accumulate over a long observation period, making the first image look completely unlike the last one. Our approach to address this problem is to use a sliding window technique inspired by data stream clustering [1] along with feature representations extracted from pre-trained CNNs [13, 27]. First, we create small sequences, that we call *snippets*, of similar images by maximizing similarities within a sliding window and then stitching these snippets to effectively capture the evolution of the objects based on overlaps and a dynamic inclusion criteria. This stitching results in sequences where images of the same sequence represent the same object (body part) and captures all stages of decomposition from fresh to skeletal. These sequences are essentially clusters of images with a

Code available at https://github.com/saramsv/SChISM.

Figure 1: Image examples of two classes, foot and hand, are shown in early (left) and late (right) decomposition stages

temporal attribute. We refer to our method SChISM as Semantic Clustering via Image Sequence Merging.

To evaluate our method, since to our knowledge there is no other work tackling the same or closely similar problem, we compared SChISM with two works that have reported to outperform other unsupervised clustering methods. First is a pre-trained CNN-based image clustering technique [13] which does not involve any model training. The second technique is DeepCluster [4] which is the state of the art clustering based on unsupervised visual representation learning. We use the general clustering metric, purity [20] for comparison. We also introduce new goodness-of-fit metrics that are more suitable for datasets with evolving content in Section 4.2.

In addition to above comparisons, we also tested SChISM on a collection of mugshot photos called MORPH [25] that contains mugshots of different individuals over time (ranging from a few days to a few years apart). This dataset has similar characteristics to the decomposition dataset, where the goal is to trace and recognize faces as they age. Results show that SChISM is capable of clustering images with 92.30% purity in the human-decomposition data whereas the pre-trained CNN-based image clustering and a trained version of DeepCluster resulted in 83.50% and 85.99% purity respectively while all three methods were set to generate the same number of clusters. In addition, clustering the mugshots using SChISM resulted in 99.88% purity for 11 subjects in 15 clusters.

In the rest of this paper, we briefly survey related work in Section 2. We then describe the details of our method in Section 3. Section 4 describes the datasets used in this work to evaluate SChISM, our evaluation technique, and results. We then conclude the paper with a conclusion in Section 5.

2. Related work

An evolving dataset, for instance images of human decomposition, can be clustered based on stages of evolution, such as the decomposition state or based on the conceptual object depicted in the image, such as the body part. Multiple Clustering methods have emerged as a result of seeking alternative clusterings that group a given dataset into clusters that exhibit different aspects of similarity. The works in [2, 24, 35] build clusters based on dissimilarity and the quality of the clusters, by forcing new clusters to be different than existing ones. In the meta-clustering method presented in [5], several alternative clusterings are found so that users can decide what set of clusters fit their need best. Similarly in [17], authors find multiple clusterings by minimizing the correlation between them through an objective function. In [6], alternative clusterings are obtained by maximizing the likelihood of each of the alternative clusterings over the data, while minimizing the similarity between them.

In all of these methods, each set of clusters is based on a single criterion. In our case however, we aim to jointly cluster the data based on two criteria: the concept of a body part and the concept of the decomposition stage.

Multi-view clustering emerged from attempts to cluster objects based on their semantic and conceptual similarities even though they might have different appearance [33, 15]. Although it might seem that we can map our problem to multi-view clustering by considering body parts as classes and different camera positions as the views, there is a fundamental difference that makes multi-view methods less suitable for the end goal of our work. In our case, clusters that include images of the same object with multiple views also evolve over time due to decomposition. Thus, the same view of the same body part appears differently depending on the stage of the decomposition.

Recent work has explored the combination of image clustering and deep representation learning [21, 29, 13]. Guérin et al. [13] studied the effect of using feature representations obtained from pre-trained CNNs on image clustering and showed that using feature representations generated using such networks results in better quality clusters.

Other works [34, 30, 4, 4] present end-to-end methods for unsupervised feature representation learning of images. Yang et al. proposed a recurrent framework for joint unsupervised learning of deep representations and image clusters by leveraging the fact that good representations are beneficial to image clustering which can be used to supervise the representation learning process [34]. In another work [30], authors trained a task-specific deep architecture for clustering. In DeepCluster, Caron et al. [4] present an end-to-end method that consists of a collaboration between clustering and classification for feature representation learning in large scale datasets in an unsupervised manner. In our work, we

utilize deep learning representations as in the above. However, these methods cannot address the implicit constraints imposed by, for example, temporally evolving objects such as the process of human decomposition alone.

Our method is partially inspired by techniques in the area of data stream clustering, which is used to monitor, for example, urban traffic and live update of stock trading. Stream clustering deals with large amounts of data that cannot be stored in memory and thus random access is not possible. Algorithms used for this purpose handle the evolution and changes in the number of clusters as new batches of data come in. One common approach in stream clustering is to use a sliding window, introduced by Aggarwal et. al [1], to keep track of how cluster centers change as new data points are streamed into the algorithm. Aggarwal et. al [1] used a sliding window instead of one-pass clustering to provide a better understanding of evolving behavior of the clusters. Several other methods [16, 36] have been built on this idea to improve the efficiency and accuracy of stream clustering.

However, the problem that we are tackling in this work cannot be directly mapped to stream clustering. In the case of evolving objects, not only does the number of clusters vary from one observation to another, but also the overall object representations change dramatically through time. However, inspired by past works, we use a sliding window along with a dynamic inclusion criteria to build a sequence of evolving images belonging to the same class.

3. Method

The goal of our method is to group large collections of unlabeled but semantically-related images. Even though semantically-related, the images may have distinct appearances due to, for example, evolution over time or representing different context. We further assume that we have partial metadata such as the timestep for each image and some implicit constraints such as the presence of images for some semantic concepts at each timestep.

First, we want to emphasize that such situations are not uncommon in cases where the image collection is subject to certain rules or protocol. Second, we would like to leverage the semantic relationships and implicit constraints to produce a fully unsupervised clustering algorithm that is capable of using these constraints to produce more accurate semantic clusters. Conceptually, our approach can be described as a penalized optimization problem where we optimize for the visual similarity of images within groups but penalize for the violation of the implicit constraints.

In the case of our dataset, the context metadata includes the subject and timestep. Each body part represents a distinct semantic concept. Decomposition makes the images of the same body part look different over time and our implicit constraints are defined by the data collection protocol that requires images of each body part at each observation. We do not, however, have body-part labels in our metadata.

While image similarity can easily group each body part into a single cluster for a specific short duration in time where the state of decomposition is the same, similarity breaks down over longer periods. To address this, we penalize sequences representing short timespans or sequences with long time gaps. In essence, we aim to minimize the loss function of the following kind:

$$loss = \frac{1}{|S|}(\sum_{k=1}^{|S|}(1 - (\sum_{j=1}^{|S_k|}\sum_{i=1}^{|S_k|} Sim(x_i, x_j))) + \frac{1}{|S_k|}) \quad (1)$$

where S denotes the generated sequences, S_k is the k-th sequence, $|S|$ is the total number of sequences and $|S_k|$ is the total number of images in the k-th sequences, and x_i is the i-th image in a given sequence. The entire parameter search space is extremely large as there are 9 different classes for which images have to be clustered per timestep and then sequenced through 50 timesteps on average for each subject, for a total of 500 subjects (million images). To make this search feasible, we break it into two stages: 1) grouping of images into short sequences (snippets) that maximizes the similarity within that snippet, and 2) stitching the snippets into longer sequences (final clusters) with minimal gaps. In the remainder of the paper we use the term snippets to refer to partially constructed semantic clusters and use the term stitching to denote the process of iteratively enlarging these incomplete semantic clusters via semantic similarity. We use this terminology to highlight the difference between the merging of images into clusters based on image similarity and based on semantic similarity as specified in the loss function above. The two terms were chosen to indicate that semantic similarity concerns the challenge of constructing contiguous time sequences of the semantic concepts.

Our method consists of three main steps, shown in Figure 2. First, we generate feature representations from input images using a CNN model pre-trained on ImageNet [8] (Section 3.1). Second, we group images into snippets (Section 3.2), and finally, we stitch similar snippets together to form long sequences (Section 3.3). In the following, the details of each step are provided.

3.1. Producing Image Features

In evolving image data such as images of human decomposition, each timestep has a group of images, subsets of which belong to various classes. We denote these images by $\text{img}_{ti} \in N$, where $t \in \{1, 2, \cdots, T\}$, $i \in \{1, 2, \cdots, m\}$ for T timesteps and m images per timestep, and N represents the set of all images. Note that the number of classes in each timestep is less than or equal to m since multiple images in each timestep may belong to the same class.

The first step in our method is to extract from each image feature representations that are used to capture image characteristics and serve as a basis for comparisons. For this,

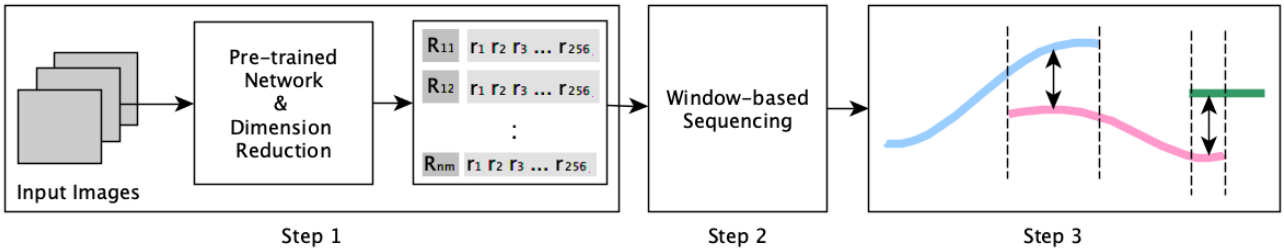


Figure 2: The overall architecture of our proposed method is shown. Step 1: Input images are mapped to feature vectors. Step 2: The neighboring feature vectors, are then compared to each other and snippets of similar images are created. We use a sliding window on the timesteps to find the neighboring feature vectors (shown in Figure 3). Step 3: Snippets are then stitched together to form longer sequences that capture the entire evolution of objects.

we feed the images into a pre-trained CNN model excluding the last fully-connected and softmax layers. In this work we used ResNet50 [14]. Other CNNs such as Inception [28] may also be used. The resulting features are then stored as feature vectors for each input image. In the case of using ResNet, each vector has a length of 2048. We denote the feature representation for image i from timestep t as R_{ti}. Inspired by Caron et al. [4], we reduce the length of these representations to 256 using Principle Component Analysis [31] to improve the overall run-time of our method.

3.2. Window-based Sequencing

In the decomposition data, typically there are one or more images representing the same class at each timestep. Additionally, the same decomposition stage may correspond to multiple consecutive timesteps and the time span may vary for different body parts. For example, the first and last 3 timesteps might represent fresh and skeletal stages respectively. As a result, there is often more image level similarity within the images of the same decomposition stage rather than across stages, which makes it a challenge to find and trace all stages of decay for a specific class without confusing it with other classes.

SChISM leverages the fact that images from neighboring timesteps are more similar to one another in terms of their local features than those from more distant timesteps. We use this constraint in our data to reduce the size of our search space and create sequences of similar images from the same class over time. Given a series of consecutive timesteps T, we define a sliding window W. Each image representation R_{ti} in W is compared to all of the images within the window except for the images of its corresponding timestep (Figure 3). If the similarity between R_{ti} and another image $R_{t'j}$ where $t' \in W$ and $t' \neq t$, is greater than a threshold, $R_{t'j}$ is added to the short sequence (snippet) that R_{ti} is a member of. If such a snippet does not exist, it is created with the two images included.

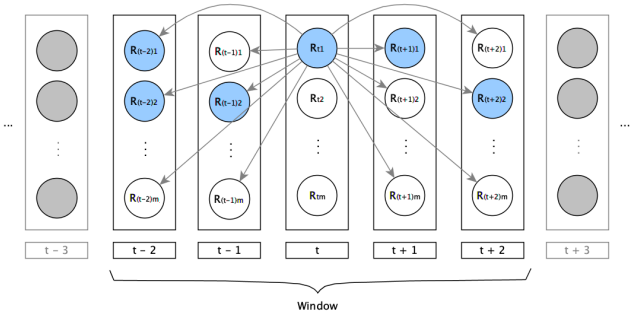


Figure 3: Each image in timestep t is compared to all images in other timesteps within the boundary of the sliding window W. After comparing all images of t, the sliding window is moved forward by one timestep.

When image classes change over time, the level of similarity between images of the same classes may vary depending on the timesteps and the state of the decomposition. Therefore, if a constant threshold is used to decide if an image should or should not be added to a snippet, classes may be miss-linked. We use a dynamic threshold to overcome the varying similarities. The threshold is set to

$$max(\alpha \times Sim_{max}(R_{ti}, R_{t'}), \beta) \tag{2}$$

where α and β are constant values. This process results in a series of snippets in which images that have the most similarity throughout time are connected, essentially grouping an image class along with its evolution. For image comparison, any two vectors, R_{ti} and $R_{t'j}$, are compared using cosine similarity as

$$Similarity(R_{ti}, R_{t'j}) = \frac{R_{ti}.R_{t'j}}{\|R_{ti}\| \, . \, \|R_{t'j}\|}. \tag{3}$$

3.3. Stitching Short Sequences

Due to the possibility of having multiple images for each class at any given timestep, the resulting snippets may have image or time overlaps. To maximize the length of final sequences for each class, we use three levels of stitching.

First, we stitch snippets that share one or more images. We call this image-overlap stitching.

Second, we stitch snippets with temporal overlaps provided that their similarity is above a constant threshold. To do so, we sort the images in each snippet based on their corresponding timesteps and find snippets that have time overlaps. For each pair of snippets, we set the one that starts with images from earlier timesteps as the first and the other one as the second. The tail of the first snippet is compared with the head of the second. The comparison is done by measuring the cosine similarity between every two image pair in the time overlap. If the average similarity of the overlap is greater than η, the two snippets are stitched together to form a longer sequence. Note that $\eta < \beta$. While this gives a second chance to stitch snippets of the same class that have not yet been grouped together using the moving window, reducing β does not have the same effect. That is because β considers inter-image similarity, while η considers inter-snippet similarity.

Third, we attempt at stitching snippets with the goal of filling the gaps that we do not expect them to have based on implicit constraints on the data, namely knowledge of the possible timesteps that exist in the data for a particular subject. To this end, each snippet is only compared against snippets of images that have time intersection with the missing timesteps in the current snippet. The comparison is done using cosine similarity between the average feature vectors of the two compared snippets. The snippets with the highest similarity are then stitched together.

4. Experimental Setup

In order to evaluate SChISM, we used images depicting human decomposition as our primary dataset as well as the MORPH dataset. Both datasets are described in Section 4.1. In section 4.2, evaluation metrics are provided. The cluster evaluation process and interface are described in Section 4.3. Finally, we present the results in Section 4.4.

4.1. Datasets

4.1.1 Human Decomposition Dataset

This image collection consists of one million photos taken of decomposing humans donated to the Forensic Anthropology Center in an 8-year period. These subjects are placed into what is known as the "Body Farm" where the different stages of decomposition are studied. The photos are taken periodically from various angles to capture different stages of body decomposition.

The images are taken daily and stored based on an ID associated with the subject and the date of the photograph. The photographer has a protocol to follow, so that all portions of the body are captured. However, due to different body placement positions and the changing of photographers over the years, the content of the photos is always changing and difficult to predict.

The main classes in this dataset are *arm, hand, leg, foot, full body, torso, backside, head, plastic* (which covers the body in some pictures), and *stake* (subject identifier). The number of images taken from the bodies for each day varies. While on average, 36 photos were taken each day, the minimum number of photos was 1 and the maximum was 358. Additionally, the number of days that each body is kept in the "Body Farm" varies for different subjects depending on how fast they decay.

4.1.2 MORPH Dataset

The MORPH dataset [25] contains mugshots collected over a span of 5 years with images of the same subject taken in real world conditions and not in controlled environments. The dataset also contains metadata in the form of age, gender, and race. The dataset has $55,134$ images of $13,618$ subjects. Having information about the age of the subjects for any given mugshot, we created a similar condition to our dataset by considering the age as the timestep concept.

4.2. Metrics

The goal of our method is to group images from the same body parts together, even-though they may look different due to decay, such that the inclusion of images of the same body part from all possible consecutive timesteps in the same cluster is maximized while the gaps in each cluster are minimized.

To evaluate the clusters produced by SChISM, we use the purity metric [20] which is defined as the ratio of correctly clustered images with respect to the dominant class in clusters, to the size of the clusters as in the following:

$$Purity_{class} = \frac{\sum_{c=1}^{\#clusters} C_c - M_c}{\sum_{c=1}^{\#clusters} C_c} \tag{4}$$

where C is the set of clusters for the given class and M is the number of misclustered images. However, because purity increases with an increase in the number of clusters, it cannot assess the quality of clusters with evolving contents alone. Therefore, we also define three new metrics namely 1) *gap*, 2) number of *essential clusters*, and 3) *inclusion*.

Gap is defined as the number of missing images corresponding to consecutive timesteps in each cluster. For example, if a subject is photographed over 10 sessions, the corresponding timesteps are $\{t_1, t_2, \cdots, t_{10}\}$. For

a given body part of the same subject if it is photographed in every session, the ideal scenario for the resulting cluster should include images corresponding to all timesteps. If the timesteps captured in the cluster are $\{0, 0, 1, 1, 0, 1, 1, 1, 1, 0\}$ (1 if there is an image corresponding to the timestep in the cluster, 0 otherwise), the gap sizes are $\{2, 1, 1\}$ (the total gap for the cluster is 4) and the size of the snippets are $\{2, 4\}$ (total length of the sequence is 6). The smaller the total gap values are the better the clusters are, in terms of tracing decomposition.

Clustering may result in multiple clusters for each class. To identify the most relevant clusters for each class, we define the **essential-cluster** metric, which is the number of non-subset clusters produced for a given class. As an example for an essential cluster, if clusters C_1 and C_2 include images for the same class corresponding to timesteps $\{t_1, t_2, t_4, t_5, t_6\}$ and $\{t_5, t_6\}$ respectively, C_1 is considered as an essential cluster rather than C_2, since C_2 is a subset of C_1. The lower the number of essential clusters for each body part (class) the better the performance of the clustering method is. In another word, the ideal scenario for each body part is to have one single cluster that includes images for all timesteps. Note that this evaluation metric can only be calculated with known class labels.

Inclusion is defined as the total number of timesteps included in the essential clusters for each body part. In other words, it indicates how many timesteps for each class are captured within essential clusters.

4.3. Cluster Evaluation

The human decomposition dataset is not labeled. In order to evaluate the performance of SChISM, we labeled a subset of the dataset to be used as test data. We developed a web interface to facilitate the labeling process and visual evaluation. Using the interface, one can label the cluster with a class name as well as selecting images that are incorrectly assigned to a given cluster with respect to the dominant class in the cluster. We used this interface to facilitate and speed up the manual labeling of our test data which include $34,476$ images corresponding to 10 randomly selected subjects from the human decomposition dataset.

4.4. Results

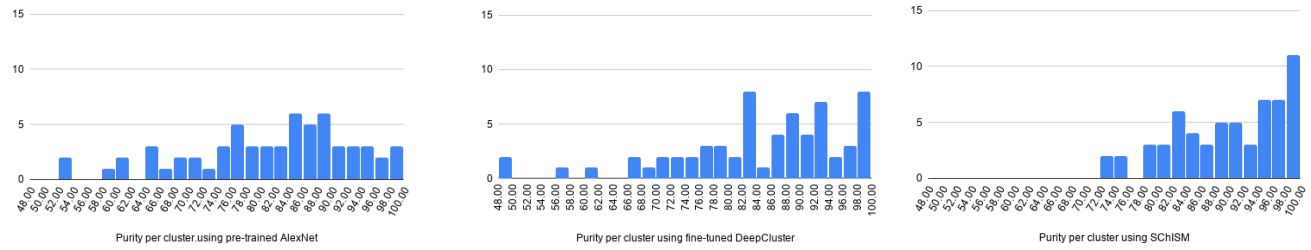
In order to test our method, we used the $34,476$ labeled images mentioned in Section 4.3. We compared the resulting clusters from SChISM with those obtained from a naive baseline as well as the following methods from [13] and [4]: a) pre-trained CNN-based image clustering, b) pre-trained DeepCluster, and c) trained DeepCluster, on the $34,476$ selected images using metrics introduced in Section 4.2 as well as the purity metric.

The **naive baseline** simply uses a non-trained CNN to map the images to feature representations and then clusters them using KMeans. **Pre-trained CNN-based image clustering** [13] is a trained version of the naive baseline method where the network is pre-trained on a common dataset such as ImageNet and then feature representations obtained from applying the network on our test data is fed to KMeans for clustering. We used ResNet50 as the CNN for both approaches for the sake of comparison with SChISM. The **pre-trained DeepCluster** method [4] consists of clustering our test data using DeepCluster pre-trained on ImageNet. Finally, we compared our results with **trained DeepCluster** which is trained on our data. Note that training DeepCluster is an unsupervised process. We did not compare against a supervised method such as training or fine-tuning a CNN on our data since our method is unsupervised and we do not have training data. The purity histograms for pre-trained CNN-based image clustering, trained DeepCluster, and SChISM are shown in Figure 4.

Table 1 shows the average purity for each class (body part) across all 10 subjects, as well as additional statistics. The number of clusters used for all methods was set equal to the number of clusters obtained from SChISM which was, on average, 72 clusters for each subject. The hyper-parameters used in our implementation of SChISM were $\alpha = 0.99, \beta = 0.7, \eta = 65$ and $W = 4$. Our analysis on different values for α, β, and η show that the higher the values, the more restrictive the inclusion criterion becomes and therefore results in a larger number of sequences and clusters and higher purity values. Lower values however, result in loosening the criterion, smaller number of sequences and clusters, and lower purity values. Higher values for W increases the chance of images from distant days being compared with each other and therefore increases computation.

Note that while increasing SChISM's hyper-parameter values results in more clusters and consequently higher purity, higher number of clusters has the same effect in other unsupervised clustering methods as well. However, Figure 4 shows that SChISM results in higher purity for the same number of clusters as used by other methods.

In addition, we further evaluated our method using the number of essential clusters, gaps, and the inclusion metrics introduced in Section 4.2. The results along with a visualized example of the clusters generated for one subject are shown in Figure 6. As Figures 6a and 6b indicate, images from similar number of timesteps are captured using smaller number of essential clusters in SChISM compared to that of pre-trained CNN-based image clustering and the trained DeepCluster. This indicates that sequences are generally longer in SChISM than in the other methods. In addition, we noticed that some classes can have zero clusters in the other methods. For example, for class 'arm', the other methods did not produce any cluster for some of the subjects. Such scenarios do not happen in SChISM due to its temporal matching. Furthermore, histograms on gaps for



(a) (b) (c)

Figure 4: Purity histogram for the clusters from pre-trained CNN-based image clustering, trained DeepCluster and SChISM. The majority of the clusters obtained using SChISM have purities higher than 84% for the same number of clusters.

Table 1: Average purity per body part as well as mean, standard deviation, median and min for all clusters with at least 5 images are provided. These values are averaged for the 10 selected subjects. The number of clusters for all methods were set to the same value obtained from SChISM for a fair comparison.

Method	**Average purity per class (%)**										**Statistics**			
	Stake	**Foot**	**Head**	**Full**	**Plastic**	**Torso**	**Arm**	**Leg**	**Back**	**Hand**	**Average Purity**	**Std**	**Med**	**Min**
Naive baseline	49.71	79.78	79.57	76.49	68.01	65.78	60.09	80.18	56.86	76.11	**72.71**	18.78	25.73	74.02
Pre-trained CNN	99.27	88.004	92.29	75.02	94.22	78.33	77.16	81.77	69.03	83.31	**83.50**	11.78	85.13	52.49
Pre-trained DeepCluster	96.57	81.34	91.69	77.29	96.11	78.53	78.05	78.14	62.41	81.93	**81.24**	14.89	83.12	31.82
Trained DeepCluster	96.85	90.73	94.07	84.97	90.89	79.91	87.92	82.87	69.93	86.76	**85.99**	11.37	88.52	49.72
SChISM	99.14	95.42	96.21	88.03	96.96	85.88	88.23	87.32	84.58	95.41	**92.30**	**7.27**	**91.93**	**72.22**

Figure 5: Example clusters obtained using SChISM.

all clusters with more than 5 members generated using pre-trained CNN-based image clustering, trained DeepCluster, and SChISM are shown in Figures 6f, 6g, and 6h and show that clusters generated using SChISM have minimum gaps compared to the other methods. We consider clusters with less than 5 members as outliers with images that do not capture body parts and could not be stitched to any of the larger sequences. We did not include the naive baseline and the pre-trained DeepCluster since pre-trained CNN-based image clustering and trained DeepCluster are the more accurate versions of the two respectively. Purity values for all approaches, however, are shown in Table 1.

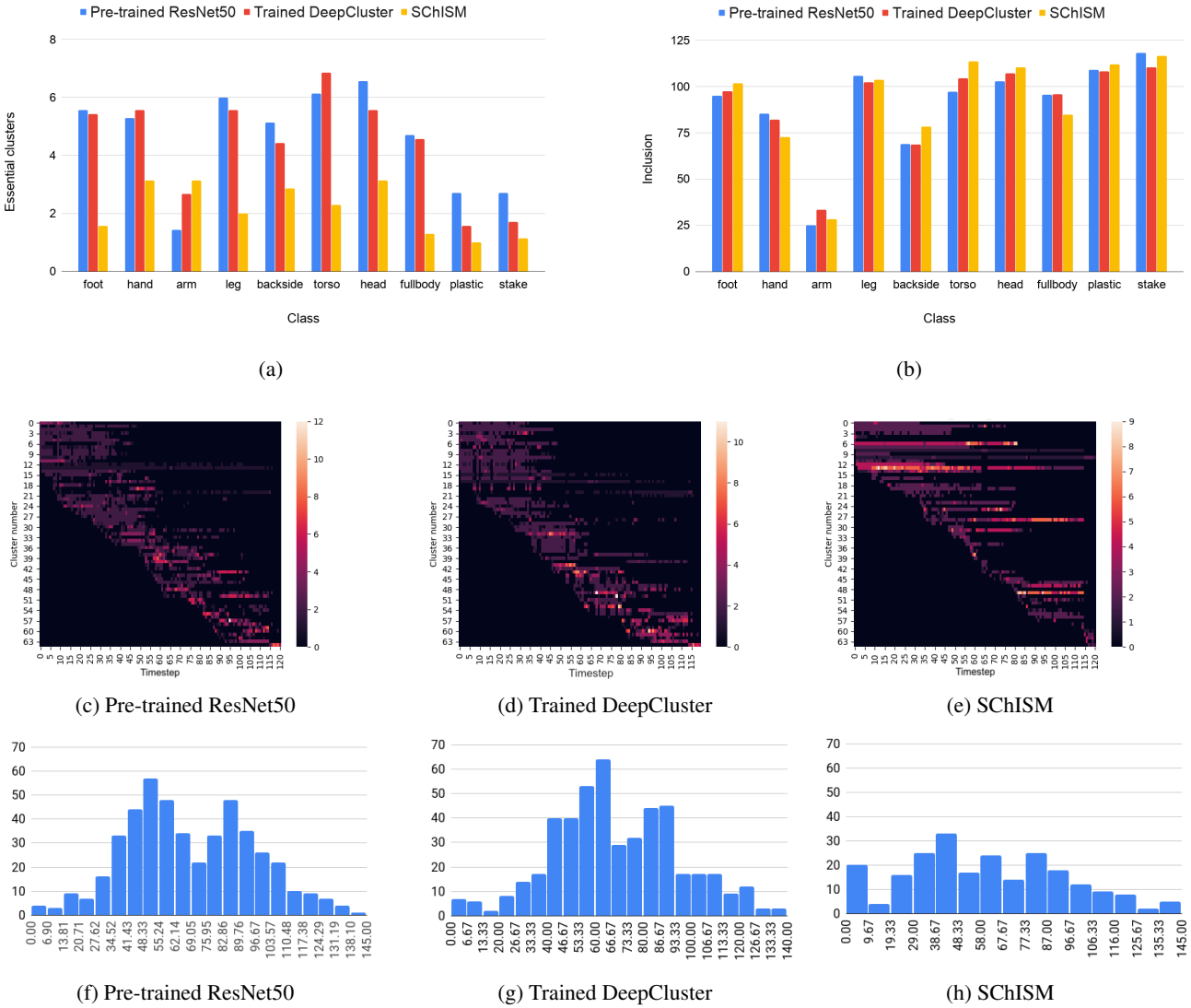
Finally, we clustered mugshot images using SChISM to assess the performance of our method on a different dataset with temporal evolving content. We selected subjects with at least 5 timesteps which resulted in 417 images from 11 subjects. We then used SChISM to group images based on subjects irrespective of their age (date of the mugshot). The result and statistics on the clusters are shown in Table 2. Figure 5 shows two clusters generated using SChISM for *foot* and a subject at ages 41, 50, 51 and 52 from the decomposition and the MORPH datasets.

5. Conclusion

Unsupervised clustering is useful for making sense of large unlabeled image collections, and can be used to accelerate manual labeling of such collections. Real-world image datasets with evolving features, however, pose challenges. We presented an unsupervised clustering technique that leverages the evolutionary characteristics and creates sequences of similar images over time using a neighboring comparison strategy with a dynamic inclusion criteria. We also introduced several metrics suitable for collections representing evolution of objects and evaluated our method on a large collection of images depicting human decomposition

Table 2: Statistics on clusters generated for the MORPH dataset using SChISM.

	#Test images	#Subjects	#Clusters	Avg. purity	Std.	Med.	Min.
MORPH	417	11	15	99.87%	0.48	100%	98.14%



(a) (b)

(c) Pre-trained ResNet50 (d) Trained DeepCluster (e) SChISM

(f) Pre-trained ResNet50 (g) Trained DeepCluster (h) SChISM

Figure 6: (a) and (b) compare pre-trained CNN-based image clustering, trained DeepCluster, and SChISM through the number of essential clusters and number of timesteps captured for each body part (inclusion). The plots show that SChISM was able to capture same or higher number of timesteps in smaller number of essential clusters. (c), (d) and (e) compare the clusters generated from the methods for one subject respectively through a visualization. Color indicates the number of images in each timestep and cluster. The plots indicate that SChISM generates clusters with longer sequences and with less gaps in timesteps compared to the other methods. (f), (g), and (h) show gap histograms for clusters and indicate that clusters generated using SChISM have minimum gaps compared to the other methods.

as well as the MORPH dataset. We further compared our method with a naive baseline, pre-trained CNN-based image clustering, pre-trained DeepCluster, and trained DeepCluster. Results show that our method produces clusters with higher purity, shorter gaps, and better inclusion for the human decomposition images compared to the other methods conditioned to produce the same number of clusters.

Acknowledgements

This work was supported by National Institute of Justice Awards 2016-DN-BX-0179 and 2018-DU-BX-0181.

References

[1] Charu C Aggarwal, Jiawei Han, Jianyong Wang, and Philip S Yu. A framework for clustering evolving data streams. In *Proceedings of the 29th international conference on Very large data bases-Volume 29*, pages 81–92. VLDB Endowment, 2003.

[2] Eric Bae and James Bailey. Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 53–62. IEEE, 2006.

[3] Miguel A Bautista, Artsiom Sanakoyeu, Ekaterina Tikhoncheva, and Bjorn Ommer. Cliquecnn: Deep unsupervised exemplar learning. In *Advances in Neural Information Processing Systems*, pages 3846–3854, 2016.

[4] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

[5] Rich Caruana, Mohamed Elhawary, Nam Nguyen, and Casey Smith. Meta clustering. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 107–118. IEEE, 2006.

[6] Xuan Hong Dang and James Bailey. Generation of alternative clusterings using the cami approach. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 118–129. SIAM, 2010.

[7] Michiel JL De Hoon, Seiya Imoto, John Nolan, and Satoru Miyano. Open source clustering software. *Bioinformatics*, 20(9):1453–1454, 2004.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia 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, 2009.

[9] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[10] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*, pages 766–774, 2014.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge 2007 (voc2007) results. 2007.

[12] Takayuki Fukui and Toshikazu Wada. Commonality preserving image-set clustering based on diverse density. In *International Symposium on Visual Computing*, pages 258–269. Springer, 2014.

[13] Joris Guérin, Olivier Gibaru, Stéphane Thiery, and Eric Nyiri. Cnn features are also great at unsupervised classification. *arXiv preprint arXiv:1707.01700*, 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Ling Huang, Hong-Yang Chao, and Chang-Dong Wang. Multi-view intact space clustering. *Pattern Recognition*, 86:344–353, 2019.

[16] Richard Hyde, Plamen Angelov, and Angus Robert MacKenzie. Fully online clustering of evolving data streams into arbitrarily shaped clusters. *Information Sciences*, 382:96–114, 2017.

[17] Prateek Jain, Raghu Meka, and Inderjit S Dhillon. Simultaneous unsupervised learning of disparate clusterings. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 1(3):195–210, 2008.

[18] Renjie Liao, Alex Schwing, Richard Zemel, and Raquel Urtasun. Learning deep parsimonious representations. In *Advances in Neural Information Processing Systems*, pages 5076–5084, 2016.

[19] Hongfu Liu, Ming Shao, Sheng Li, and Yun Fu. Infinite ensemble for image clustering. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1745–1754, 2016.

[20] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.

[21] Sara Mousavi, Dylan Lee, Tatianna Griffin, Dawnie Steadman, and Audris Mockus. An analytical workflow for clustering forensic images. *arXiv preprint arXiv:2001.05845*, 2019.

[22] Sara Mousavi, Dylan Lee, Tatianna Griffin, Dawnie Steadman, and Audris Mockus. Collaborative learning of semi-supervised clustering and classification for labeling uncurated data. *arXiv preprint arXiv:2003.04261*, 2020.

[23] Sara Mousavi, Ramin Nabati, Megan Kleeschulte, Dawnie Steadman, and Audris Mockus. Machine-assisted annotation of forensic imagery. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1595–1599. IEEE, 2019.

[24] Donglin Niu, Jennifer G Dy, and Michael I Jordan. Multiple non-redundant spectral clustering views. 2010.

[25] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345. IEEE, 2006.

[26] Gerald Schaefer and Michal Stich. Ucid: An uncompressed color image database. In *Storage and Retrieval Methods and Applications for Multimedia 2004*, volume 5307, pages 472–480. International Society for Optics and Photonics, 2003.

[27] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

[28] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[29] Fei Tian, Bin Gao, Qing Cui, Enhong Chen, and Tie-Yan Liu. Learning deep representations for graph clustering. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[30] Zhangyang Wang, Shiyu Chang, Jiayu Zhou, Meng Wang, and Thomas S Huang. Learning a task-specific deep architecture for clustering. In *Proceedings of the 2016 SIAM*

International Conference on Data Mining, pages 369–377. SIAM, 2016.

[31] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[32] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487, 2016.

[33] Yu-Meng Xu, Chang-Dong Wang, and Jian-Huang Lai. Weighted multi-view clustering with feature selection. *Pattern Recognition*, 53:25–35, 2016.

[34] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5147–5156, 2016.

[35] Sen Yang and Lijun Zhang. Non-redundant multiple clustering by nonnegative matrix factorization. *Machine Learning*, 106(5):695–712, 2017.

[36] Aoying Zhou, Feng Cao, Weining Qian, and Cheqing Jin. Tracking clusters in evolving data streams over sliding windows. *Knowledge and Information Systems*, 15(2):181–214, 2008.