Video-based Human Activity Analysis: An Operator-based Approach

Xiao Bian
North Carolina State University
Department of Electrical and Computer Engineering
27606, Raleigh, NC, USA
xbian@ncsu.edu

Hamid Krim
North Carolina State University
Department of Electrical and Computer Engineering
27606, Raleigh, NC, USA
ahk@ncsu.edu

ABSTRACT

Human activity in video sequences may be viewed as a sampled trajectory on a low dimensional manifold embedded in a high dimensional ambient space. Due to the unknown underlying manifold structure of the image frames, we propose a novel framework to define a neighborhood for high dimensional data which, when acted upon by a mapping operator, results in a subset in an a priorily well defined range space. We exploit the so-called correlation filtering with a specifically selected output response to effectively approximate the data manifold by way of encoding local neighborhoods on it. This helps us propose an unsupervised learning algorithm of human activity, and demonstrate its performance in classifying and clustering of different activities taking place in observed video sequences.

Keywords

Manifold learning, Human activity, Correlation filter

1 INTRODUCTION

Video-based human activity analysis plays an important role in video content analysis such as video surveillance and video indexing. Its inherent complexity is due to the high dimensionality and nonlinearity of the associated feature space. Using the low dimensionality which is intrinsic to the human activity dynamics, numerous manifold learning techniques have been proposed [AWAR11] [BR07] [SKN12]. By reducing the dimension of the feature space, one can reduce the impact of 'the curse of dimensionality', and thereby potentially improve the human activity classification performance [BNS06]. The raw data on its own, is, however, insufficient to reflect the necessary information for characterizing the underlying manifold structure (or associated tangent space) of the frame sequences in various human activities. This problem was overcome by shape-based methods [SKN12] as well as in [GBS+07] by exploiting the well defined shape manifold of the silhouette contours in images of video sequences. These unfortunately, exhibit limitations when the targets (human images in the frames) undergo a topological change, or when there is more than one target in a scenario to be analyzed.

The key idea underlying manifold learning is the assumed ability to faithfully encode its essential local information . This local information is, in turn, determined by the k-nearest neighbor technique. It is thus critical to find a robust way to identify the neighborhood for each data sample (i.e. frame). In this paper, we choose to define a neighborhood in a high dimensional space by way of a sequence of correlation operators defined on the manifold. The manifold structure is thus now encoded in the operator sequences which act on the corresponding neighborhoods. We also exploit the correlation filters [BB10] to develop an unsupervised learning algorithm for capturing human activity characteristics under the proposed framework, and to thereby demonstrate its learning capacity by interpolation and activity transition detection. The clustering and classification potentials of the algorithm are in addition illustrated, and their performance demonstrated.

The remainder of this paper is organized as follows, in Section 2, we describe the mathematical formulation of the local information encoding by way of an associated operator sequence on the underlying data manifold of interest. In Section 3, we introduce the detailed algorithms used in analyzing the human activity video sequences of interest. In

Section 4, we substantiate our development by running clustering and classification experiments on real data, and demonstrate the performance of our proposed method. In Section 5, we conclude with some remarks and with a brief discussion of future research directions.

2 CLUSTERING OPERATORS ON MANIFOLD

It is a clear fact in human activity analysis or in general video sequence analysis, that we neither have the explicit form of the underlying manifold of the images/frames, nor do we have the ability to approximate the tangent space of this manifold. Many learning techniques have explored underlying structures of manifolds by tangent space estimation, which in turn is approximated by neighborhood points determined by k-nearest neighbors [Row00] [Don03] [BN01]. However, distance measures become very sensitive with the increase of dimensionality, and the proportional difference between a farthest point and a nearest point vanishes, making the k-nearest neighbor approach unreliable [HK00] [HK10]. In addition, sample points (in our case our video frames) are not necessarily uniformly distributed in ambient space making the tangent space estimation even more challenging, and learning the manifold more elusive.

When considering human activity, we may, however, assume a low dimensional structure for a given video sequence, or as a time sequence of sample points in high dimension, which may in turn be viewed as a trajectory curve on a lower dimensional manifold embedded in high dimensional space.

As discussed in the next sections, we take advantage of mapping operators to more precisely define sequential neighbors in a high dimensional space.

2.1 Bounded operators on a manifold

Building on [Yao98] [RV07], we proceed to determine "neighboring frames" by way of mapping operators applied to data to achieve a predefined desired output.

To be precise, the following definition of neighbors is in order,

Definition 1 *Consider a given continuous mapping $\mathscr{T} : U \mapsto V \subset R^n$, U is a subset of a data manifold M which is embedded in R^n. Given $\sigma > 0$. **A σ-neighborhood** of x_0 is a subset $N \subset U$, $\forall x \in N$ satisfying $\|\mathscr{T}x_0 - \mathscr{T}x\| \leq \sigma$.*

When searching for neighbors of $x_1, \|\mathscr{T}x_1 - \mathscr{T}x_2\|$ may also be interpreted as $h : U \mapsto \mathbb{R}^+, h(x) = \|\mathscr{T}x_1 - \mathscr{T}x\|, x$ is a neighbor of x_1 if and only if $h(x) \leq \sigma$. From the theory of bounded operators [Kre78], we may alternatively provide a more general scope by the following,

Definition 2 *Consider $h : U \mapsto V$ a continuous mapping on $\mathscr{U}$, x is a **W-neighbor** of y under h, if $h(x) \in W, h(y) \in W, W \subset V$, $S = h^{-1}(W) \subset U$ is the **W-neighborhood** under h.*

The complexity and the nonlinearity of a human activity require that an operator sequence $\tilde{h} = \{h_1, \ldots, h_q\}$ be applied to cover all the degrees of freedom intrinsic to high dimensional video sequences. Each operator h_i in the sequence $\tilde{h}$ defines a neighborhood S_i, which can be seen as an approximation of a tangent space of the data manifold. The set of $S = \{S_i\}$ covers the whole high dimensional space of video sequences of a given human activity. Practically, this is also a way to overcome the difficulties of k-nearest neighbor estimation, and to encode local information of a manifold simultaneously, as noted earlier in Section 2. This is a more general and flexible depiction/representation of a local open set on a manifold than that of a local tangent space often used in manifold learning algorithms. To sum up the idea above, we may succinctly state the following,

Definition 3 *Consider a curve $\mathscr{X} : [0,1] \mapsto U \subset R^n$, given a set of operators $\tilde{h} = \{h_1, \ldots, h_q\}, h_i : U \mapsto V$, $\mathscr{X}$ is said to be covered by a W-neighborhood under $\tilde{h}$, if $\forall t \in [0,1], \exists i \in \{1, \ldots, q\}, h_i(\mathscr{X}(t)) \in W, W \subset V$.*

2.2 Representative Operators of a High Dimensional Image Space

To exploit the framework in Section 2.1 in addressing a video sequence modeling, we need to construct an operator set associated to video sequences in the class of human activities of interest. An important property of any resulting operator, is that it must be least sensitive to noise commonly encountered in measured images. Instead of vectorizing each image in R^n, and encoding the local information by selecting k-nearest neighbors in the metric space R^n, as is commonly practiced in manifold learning techniques [Row00] [Law05], we choose to radically depart from this idea. This is primarily due to the high sensitivity of Euclidean distance in R^n to common noise terms in image processing (i.e. White noise), and to the limitation brought on by "the curse of dimensionality" [HK00] [HK10]. Inspired by the successful application of filters in image processing [BB10], we proceed to choose our fore-described mapping operators from the class of 2-dimensional convolution kernels: $h_i(x) = h_i * x$. In contrast to the distance-based low dimensional subspace representation of vectorized frames, the inherent relations among

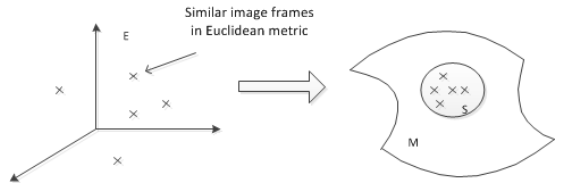

Figure 1: Similar frames may not close to each other in Euclidean metrics. The Trained Operators can map similar frames close to each other in the range space

video frames will be preserved in the neighborhood under the mappings, and the noise largely filtered out.

To more clearly define the data neighborhood, we proceed to define the covering of $X(t)$ (frame/video sequence) by an operator sequence,

Definition 4 *$h : U \mapsto V$ is a convolution operator on $\mathscr{U} \subset R^n$, then x is a neighbor of y if $h * f \in W, h * y \in W, W \subset U$.*[1]

Definition 5 *$\mathscr{X} : [0,1] \mapsto U \subset R^n$, is covered by $h = \{h_1, \ldots, h_q\}, h_i : U \mapsto V$ if $\forall t \in [0,1], \exists i \in \{1, \ldots, q\}, h_i * \mathscr{X}(t) \in W, W \subset V$.*

3 CONVOLUTIONAL OPERATOR SEQUENCE CONSTRUCTION

Convolutional kernels have been successfully used in video analysis for correlation filter-based tracking [BB10]. Trained correlation filters can robustly track targets in a video sequence. It is a special case of the filters in Definition 5 since one filter is sufficient to learn from the data sequence, and to track the target with small perturbation. In human activity analysis, a target may dramatically change because of articulation deformation. This is hence tantamount to saying that a set of operators will be required to capture all the potential information embedded in a video sequence. Building on MOSSE filters [BB10], the training of operators is carried out on the basis of a sample point set of a high dimensional curve $\mathscr{X}$ per Definition 5. For clarity as well as computational efficiency, we discuss all implementational issues in the Fourier domain. Let $F = \mathscr{F}(f)$ denote the 2D Fourier transform of a given image frame, and $H = \mathscr{F}(h)$ that of the h. In that light, and given a data set F_i, the optimal filter H^* is obtained by

$$\min_{H^*} \sum_i |F_i \bigodot H^* - G_i|, \tag{1}$$

where $G_i = \mathscr{F}(g_i)$ denotes the Fourier transform of the ideally desired output g_i, a spatial ($2D$) gaussian signal, and $\bigodot$ denotes a Schur product (i.e. an element-wise multiplication).

As shown in [BB10], the optimal solution of Eq.(1) is

$$H^* = \frac{\sum_i G_i \bigodot F_i}{\sum_i F_i \bigodot F_i^*}. \tag{2}$$

3.1 Neighborhoods Information Encoding

As noted earlier, the operator sequence with their associated neighborhood encode the information of the data manifold, and hence implicity describe a corresponding manifold of operators which captures and reflects the information of the initial data (video sequence). It is hence important that the defined neighborhood be preserved following the mapping (i.e. close points in the initial manifold space should remain close in the range space upon mapping) and be robust to noise. To this end, we also adopt a Peak-Sidelobe Ratio (PSR) as the optimization criterion of choice [BB10],

$$PSR = \frac{g_{max} - \mu_s}{\sigma_s}, \tag{3}$$

where g_{max} is the maximum value of the response; μ_s and σ_s respectively are the peak value and variance of a sidelobe.

Given $g_i = f_i * h$, f_i is the input data and h is the operator, then PSR is a function defined on the output image domain, $PSR(g_i) \in R^+$.

More specifically, neighbors of a given frame under a correlation filter operator is defined as,

Definition 6 *For $x, y \in \mathscr{U}$, h is a operator on $\mathscr{U} \subset R^n$, y is called a neighbor of x if $h * x \in W$ and $h * y \in W, W \subset U$. Furthermore, $h * x \in W \Leftrightarrow PSR(h * x) > \eta > 0$*

3.2 Learning an Operator Kernel Sequence from Data

As discussed above, we propose a specific algorithm to automatically explore the structure of the data manifold by way of a sequence of kernel convolutional operators whose manifold is more easily characterized. The cardinality of the set of operators being unknown a priori, we first randomly pick a frame f_i from the available data set. We subsequently select all points of the neighborhood of an f_i-trained operator to further train the operator H_j. Excluding all data points from the neighborhood under H_j, we pick the farthest point in the data set from S, to iteratively determine the next operator and until all data in the training set is covered by the operator sequence $H = \{H_i\}$.

The algorithm to construct an operator sequence is described below.

[1] x and y are samples/frames of a sequence.

- Given image sequences $\{f_i\}$ as a training set T
- Randomly pick one frame f from the training set T
- While T is not an empty set

  1. Train operator H from frame f. Find all frames f_{H_i} which belong to S_H(The neighborhood under H)
  2. Train operator $\hat{H}$ from S_H. Find all frames $f_{\hat{H}_i}$ which belong to $S_{\hat{H}}$(The neighborhood under $\hat{H}$)
  3. Define T as the complement of T with respect to $S_{\hat{H}}$, $T = T - S_{\hat{H}}$
  4. Define f_i to be the 'farthest' point to $S_{\hat{H}}$ in T: $f_i = \arg\min_{f_j \in T} PSR(h(f_j))$

4 VALIDATION AND EXPERIMENTS

We next carry out a series of experiments to demonstrate the performance of our algorithms for human activity analysis. We use the database in [GBS+07], which includes 9 different people performing 9 different activities individually, such as jumping, running and walking, etc. The relatively low resolution video clips with a naive foreground extraction, provide a good approximation to the noisy and imperfect environment of video surveillance. And since the textural information for each person does not impact the activity itself, we exclude the textures, and only use distance maps for each frame.

4.1 Interpolation and Segmentation

With an operator set $h = \{h_i\}$ and video frames $f = \{f_j\}$, each f_j has a index i_j from its corresponding operator h_{i_j}. The sequence of indices demonstrates the stages of a video sequence. Fig. 2 gives an example of the segmentation of a video sequence.

With the information of the operator sequence, we can interpolate between two frames, as shown in Fig. 3 and Fig. 4. Note that if data on the manifold as defined by our operator sequence is used, the interpolation should then successfully recover the true activity frame.

4.2 Activity change detection

Activity change will influence the response of operator sequences. Since the operator set is specifically designed to cover all variations in each activity, the reduced response always means the variation of activity, as shown in Fig. 5

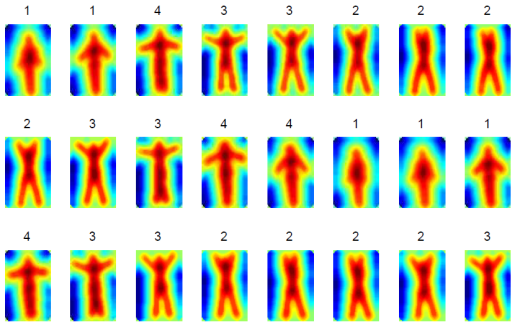

Figure 2: Segmentation of human activity based on neighborhood under each operator. The number on each frame represents the index of the associated operator

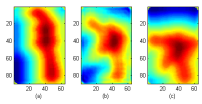

Figure 3: Human activity sequences (bending) interpolation. (b) The interpolated gesture between left and right frames

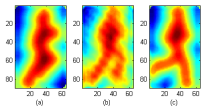

Figure 4: Human activity sequences (running) interpolation. (b) The interpolated gesture between left and right frames

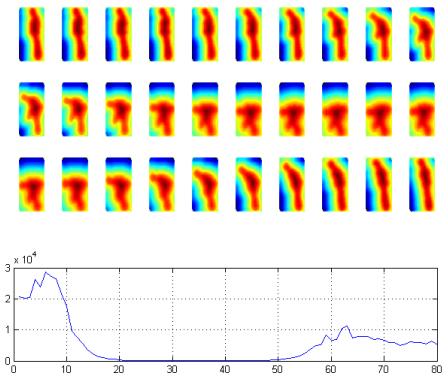

Figure 5: The response of the operator set is severely decreased with activity transition. x axis represents frame number and y axis represents PSR value

4.3 Activity Clustering

Similarity between two sequences $f^i = \{f^i_k\}$ and $f^j = \{f^j_k\}$ is defined as

$$A_{ij} = \left(\frac{N_{ij}}{N_{ii}} + \frac{N_{ij}}{N_{jj}}\right)/2 \tag{4}$$

N_{ij} = Cardinality of $S_j \bigcap f^i$. S_j is the neighborhood of f^j under their own operator set. Notice that here no further registration or alignment is needed for this metric.

In this experiment we applied a regular spectral clustering algorithm [NJW01] on the similarity matrix(Fig. 6)of all 81 video sequences. With 7 out of

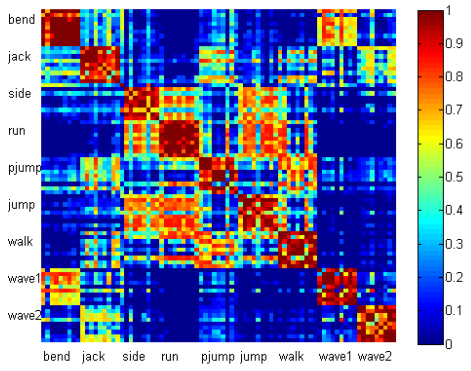

Figure 6: Similarity matrix of 9 classes of human activities, 9 realizations for each activity. 1 means the given two sequences are similar to each other and 0 vice versa

81 misclassified, 91.36% sequences are correctly clustered. The success of clustering here demonstrates that the similarity measure reflects the intrinsic closeness of different human activity sequences.

4.4 Activity Classification

We separate the database into a training set and a test set by randomly selecting 4 sequences from each activity and let the remaining 5 sequences of each activity as a test set. For an input sequence, each operator set H_i will convolve the signal and we can have a set of sequences of PSR value. We calculate the maximum PSR at each frame for each operator set, and use them as similarity features between input and the corresponding class. The input sequence is then assigned to the class of operator with maximum median, such as Id of Input = Id of $\max median(PSR(g))$

In spite of selecting relatively a small training set, the rank-1 recognition rate is 90.06%, with 31 misclassified in 315 tests.

5 CONCLUSION

In this paper, we present a novel framework of manifold learning by using operators on a manifold and propose an unsupervised learning algorithm to represent human activity sequences with a small number of operators. By using the set of operators for each human activity, we demonstrate the high performance for clustering and classification. Combined with successfully interpolating frames among sequences, the experiments show that the feature extracting by an operator set construction is fast, accurate and robust. Furthermore, from the experimental results, a high dimensional neighborhood in our framework is more robust compare to Euclidean distance, and shows potential to overcome the curse of dimensionality.

6 REFERENCES

[AWAR11] M. Abdelkader, A. Wael, S. Anuj, and C. Rama. Silhouette-based gesture and action recognition via modeling trajectories on riemannian shape manifolds. *Computer Vision and Image Understanding*, 115(3):439–455, March 2011.

[BB10] D. Bolme and Jr Beveridge. Visual object tracking using adaptive correlation filters. *Computer Vision and Pattern Recognition (CVPR)*, 2010.

[BN01] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14*, pages 585–591. MIT Press, 2001.

[BNS06] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold Regularization : A Geometric Framework for Learning from Labeled and Unlabeled Examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[BR07] J. Blackburn and E. Ribeiro. Human motion recognition using Isomap and dynamic time warping. In *Proceedings of the 2nd conference on Human motion: understanding, modeling, capture and animation*, pages 285–298. Springer-Verlag, 2007.

[Don03] D. Donoho. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Academy of Sciences of the United*, (650):1–15, 2003.

[GBS+07] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE transactions on pattern analysis and machine intelligence*, 29(12):2247–53, December 2007.

[HK00] A. Hinneburg and D. Keim. What is the nearest neighbor in high dimensional spaces. In *Proc. VLDB*, 2000.

[HK10] M. Houle and H. Kriegel. Can shared-neighbor distances defeat the curse of dimensionality. In *22nd International conference on scientific and statistical database manangement*, 2010.

[Kre78] E. Kreyszig. *Introductory Functional Analysis with Application*. Wiley, 1978.

[Law05] C. Lawrence. Algorithms for manifold learning. *University of California, San Diego, Tech. Rep. CS2008*, 2005.

[NJW01] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information*

Processing Systems, pages 849–856. MIT Press, 2001.

[Row00] S. T. Roweis. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, December 2000.

[RV07] J. Rowe and M. Vose. Neighborhood graphs and symmetric genetic operators. *Proceeding FOGA'07 Proceedings of the 9th international conference on Foundations of genetic algorithms*, pages 110–122, 2007.

[SKN12] Y. Sheng, H. Krim, and L. Norris. Human Activity Modeling as Brownian Motion on Shape Manifold. *Scale Space and Variational Methods in Computer Vision*, pages 628–639, 2012.

[Yao98] Y. Yao. Relational interpretations of neighborhood operators and rough set approximation operators. *Information sciences*, 111(1):239–259, 1998.