Orthogonal Hankel Subspaces for Applications in Gesture Recognition

Bernardo B. Gatto
Federal University of Amazonas
Manaus, Amazonas, Brazil
bernardo@icomp.ufam.edu.br

Eulanda M. dos Santos
Federal University of Amazonas
Manaus, Amazonas, Brazil
emsantos@icomp.ufam.edu.br

Waldir S. S. Júnior
Federal University of Amazonas
Manaus, Amazonas, Brazil
waldirjr@ufam.edu.br

***Abstract*—Gesture recognition is an important research area in video analysis and computer vision. Gesture recognition systems include several advantages, such as the interaction with machines without needing additional external devices. Moreover, gesture recognition involves many challenges, as the distribution of a specific gesture largely varies depending on viewpoints due to its multiple joint structures. In this paper, We present a novel framework for gesture recognition. The novelty of the proposed framework lies in three aspects: first, we propose a new gesture representation based on a compact trajectory matrix, which preserves spatial and temporal information. We understand that not all images of a gesture video are useful for the recognition task, therefore it is necessary to create a method where it is possible to detect the images that do not contribute to the recognition task, decreasing the computational cost of the overall framework. Second, we represent this compact trajectory matrix as a subspace, achieving discriminative information, as the trajectory matrices obtained from different gestures generate dissimilar clusters in a low dimension space. Finally, we introduce an automatic procedure to infer the optimal dimension of each gesture subspace. We show that our compact representation presents practical and theoretical advantages, such as compact representation and low computational requirements. We demonstrate the advantages of the proposed method by experimentation employing Cambridge gesture and Human-Computer Interaction datasets.**

I. Introduction

Gesture recognition is a very attractive research area in the computer vision field, because it provides the means to interact with machines without the need of additional devices. Gesture recognition is widely employed in human-computer interaction applications by using a predefined set of human joint motions. In order to develop such systems, in which the recognition gestures can be used to transmit significant information or controlling virtual environments, a high efficient gesture recognition framework is required. Accordingly, gesture recognition exploits high sophisticated machine learning techniques in order to work accurately in various domains.

Controlling machines employing gesture recognition is useful. However, it includes many difficulties, for instance the distribution of a gesture largely varies depending on viewpoints due to its multiple joint structures. Further, recognition and estimation of the gesture are very difficult because masked or occluded regions are often produced, requiring a robust framework. In addition, camera position, illumination conditions and pose may also increase the application overall complexity. In order to solve these problems, several methods have been introduced for gesture recognition including discriminative canonical correlation analysis (DCC) [1], [2], hidden Markov models (HMM) [3], orientation histograms [4], color based-models [5], dynamic time Warping (DTW) [6], and silhouette geometry-based models [7].

Moreover, there are applications that use external hardware to improve the recognition performance. In [8], Kinect^{TM} sensor measures depth information in order to decrease the complexity of segmenting the gesture joints, improving significantly the overall application performance. Another possibility is to employ more sophisticated devices; such as gloves with accelerometers [9] or Leap Motion Controller^{TM} [10]. Although these methods have shown high performance, in our proposed method, we are interested in employing only machine learning techniques on raw images, without making use of external hardware nor pre-processing techniques, as we understand that such devices may increase the cost of the system both computationally and economically.

In general, gesture recognition frameworks make use of pre-processing techniques in order to extract relevant information for classification. A variety of methods that employ pre-processing techniques has been proposed for gesture recognition. For instance, HOG [11], SIFT [12], SURF [13] and LBP [14] can perform efficient feature extraction. However, this pre-processing may increase the framework complexity, preventing its application on environments where there is restricted hardware. On the other side, some solutions do not require pre-processing techniques. For instance, subspace-based methods have the advantage of working directly on the raw images, as they do not require feature extraction in order to represent the image set distributions.

Instead of employing feature extraction, subspace-based methods work directly on the pixel level of the images, creating a very light hence efficient representation. In [15], it has been argued that this representation is very powerful, since the subspace of an image-set generates distinctive clusters in a low dimensional space. Based on this observation, our assumption is that we can also represent gesture images as subspaces by employing Hankel matrix in order to efficiently create the covariance matrix. Subspace-based methods have been employed in several computer vision applications, including

face recognition [16], [17], object recognition [18], [19] and hand shape recognition [20], [21].

Despite the fact that subspace-based methods can achieve high performance when applied to image set recognition, subspace-based methods are not able to cope with temporal information, as required for an efficient gesture representation. To solve this problem, we propose a new method based on clustering and sample selection in order to reduce its computational complexity and simultaneously preserving the temporal information. This new representation is mainly based on Hankel matrix formulation, where the image patterns can be stored in a manner where the ordering of the images is preserved. In this approach, we select representative samples from each image gesture set to compound its corresponding Hankel matrix. By exploiting this strategy, we obtain a smaller covariance matrix, compared to traditional methods, where we can easily extract its basis vectors.

In general, subspace-based methods ignore the intrinsic dimension of each class distribution, treating all of them as the same dimension. This leads to several problems, such as vanishing of discriminative and representative features. For instance, we can infer that different distributions have different accumulated energy in each eigenvector. Some classes may have a high compactness ratio in only the first 4 eigenvectors, for instance, achieving a very efficient representation. However, some classes may have a high spread ratio energy over its eigenvalues, where only 4 eigenvectors are not sufficient to represent such classes. Therefore, we also propose an automatic method to weight the basis vectors of each image class, in order to better preserve its intrinsic dimension.

The contributions of this study to the literature are: (1) A novel framework for gesture recognition, with no preprocessing techniques, requiring low computational resources. (2) A new representation for gesture recognition, where the samples are dynamically selected, creating a very compact representation. In addition, by employing the Hankel matrix, this new representation is able to preserve temporal information. (3) An automatic approach for basis vector weighting based on the accumulated energy strategy. In this solution, we employ all the basis vectors available for classification, without parameter tuning.

This paper is organized as follows. Section 2 describes related work on gesture recognition and image-set classification by subspace-based methods. Section 3 introduces Orthogonal Hankel Hankel Subspace Method (OHSM) by using a compact Hankel matrix representation for gesture recognition and the automatic soft weight process for subspaces. Section 4 shows the experimental results. Finally, Section 5 presents conclusions.

II. Related Work

In this section, we briefly describe the main variants of subspace-based methods, as well as the differences among these methods. Then, we also present relevant work on Hankel matrices for image-set based pattern recognition. Finally, we enumerate the main techniques employed to measure the similarity between subspaces. This description is fundamental in order to clarify the differences and improvements when comparing the proposed gesture recognition framework to current methods.

As mentioned before, most of the subspace-based methods employ discrete Karhunen-Loève transform (KLT), also known as principal component analysis (PCA), in order to generate the subspaces. These techniques are preferred because they are optimal to achieve a subspace that minimizes the mean square error.

The main advantage of using KLT subspace to represent image sets lies in its compact representation, hence, decreasing the overall classification system. This classification is usually achieved by using multiple canonical angles [22]. Recently, a variant of subspace method was introduced [21]. In this variant, called generalized difference subspace (GDS), the image set patterns are also represented as subspaces, however, the relationship between the patterns are taken into consideration by employing the concept of generalized difference between the subspaces. This algebraic formulation provides a novel discriminative transformation, where the projected subspaces produce higher recognition results compared to conventional subspace-based methods.

In despite of its high recognition results [23], [24], GDS formulation is not adequate for more advanced systems, wherein temporal structures should be classified. For instance, when the order of the patterns plays an important role in the classification system, GDS tends to decrease its performance, as will be demonstrated by experimental results in Section IV.

Hankel matrices employed to preserve temporal information is not novel. Several approaches have been introduced in order to retain temporal information to represent activity [25], emotions [26] and group activity recognition [27].

For activity recognition, the concept of Hankelets [25] has been proposed. In Hankelets, features are extracted by using a bag of features (BoF) approach to recognize activities across different viewpoints. In this method, Hankelets produce a novel representation for activities where viewpoint invariance is taken into account, keeping the activities dynamic, instead of spatial gradient information. The advantages of this method include that Hankelets are straightforward to obtain and do not require prior 3D models, camera calibration, persistent tracking or spatial feature matching.

Hankel matrices have been employed to efficiently represent spatial and temporal information in group activity recognition [27]. In this work, the problem of recognizing the interactions and the group activity from wearable cameras, such as Google Glass, is investigated. The solution arises from the combination of the temporally synchronized videos from different wearers, where Hankel matrices and movement pattern histograms are employed for feature representation.

The main concern about employing Hankel matrices to represent image sequences is that its computational cost required to efficiently extract the basis vectors is very high. In order to solve this issue, we use a strategy to decrease the number of employed images from the images sequences. We achieve

this subset by using a clustering approach and, therefore, we can construct a more compact Hankel matrix. We show by experiments that this compact representation achieves higher recognition rate when compared to the usual approach and it is computational more efficient.

The topic of selecting and weighting the basis vectors of a subspace have been investigated in the literature. For instance, in [28], the criteria of accumulated energy is employed to select the basis vectors that will represent an image-set. It is well known that the eigenvectors associated with the higher eigenvalues preserves most of the energy contained in an image-set. Therefore, selecting the first eigenvectors corresponding to 90% of the accumulated energy is a straightforward strategy that may achieve good results, without delving in a brute force parameter search.

Weighting the basis vectors of the subspace is another alternative to optimize the use of the eigenvectors. For instance, in [29], a weighed strategy is adopted to accomplish an efficient framework for face reconstruction and classification. In this work, it is observed that not all combinations of the basis vectors form a meaningful face, therefore, certain restrictions should be adopted. The weighting strategy ensures that the similarity between two subspaces is obtained at points that actually correspond to faces of the respective classes.

Generalized mutual subspace method (gMSM) [18] employs all the basis vectors in order to represent its subspace. Traditionally, the importance of the eigenvalues is taken in a binary decision, where most of the eigenvectors are discarded according to its eigenvalues. On the other hand, in gMSM, all the eigenvectors are used according to its eigenvalues in a weighed scheme, therefore, even the smaller eigenvalues contribute to the subspace, but in a lower proportion.

III. Proposed Method

In this section, first we describe the problem of gesture recognition from image sets. Next, we explain the applications of Hankel matrix ordered image set representation. After that, we introduce the procedure of creating Hankel subspaces. Then we show the procedure to select the samples in order to improve the processing time to extract the basis vectors of a given Hankel matrix. We introduce the dynamic soft weights and its advantages over the conventional method. Finally, we describe the procedure to match two Hankel subspaces to compute its similarity. Figure 1 shows the conceptual diagram of the proposed method.

A. Problem Formulation

Given a set of gesture images, which are given by $\mathbf{A} = \{\mathbf{A}_i\}_{i=1}^{M}$, where $\mathbf{A}_i$ is an image. Then, we define that $\mathbf{A}$ is ordered, so $\mathbf{A}_1 \preceq \mathbf{A}_2 \preceq \mathbf{A}_3 \preceq \ldots \preceq \mathbf{A}_M$. Then, we assume that there is a linear mapping that represents $\mathbf{A}$ set in terms of its variance, preserving its spatial and temporal information. This linear transformation is in such way that the M gesture images are converted into $k-$dimensional orthonormal vectors ordered by its accumulated energy. This new representation, $\mathbf{\Phi_A} = \{\phi_i\}_{i=1}^{k}$, provides a more compact manner to represent set $\mathbf{A}$ and its computational classification cost is therefore, greatly reduced. The Φ_A set spans a reference subspace $\mathbf{P_A}$. In literature, $k \ll M$, where discriminative information may be lost. In our proposed method, $k = M$, as all the obtained basis vectors will be employed to create a subspace and a weight will be assigned to each basis vectors ϕ_i regarding its variance. Finally, for a given gesture image set $\mathbf{Y} = \{\mathbf{Y}_i\}_{i=1}^{N}$, where $\mathbf{Y}_1 \preceq \mathbf{Y}_2 \preceq \mathbf{Y}_3 \preceq \ldots \preceq \mathbf{Y}_N$, the task is to compute a subspace $\mathbf{Q_Y}$ that represents $\mathbf{Y}$ in terms of its variance, preserving its spatial and temporal information and calculate how similar $\mathbf{Q_Y}$ and $\mathbf{P_A}$ are.

B. Hankel Matrix-based Gesture Representation

Let $\mathbf{A} = \{\mathbf{A}_i\}_{i=1}^{M}$ representing a gesture that is handled as a time series of vectors, where $\mathbf{A}_1 \preceq \mathbf{A}_2 \preceq \mathbf{A}_3 \preceq \ldots \preceq \mathbf{A}_M$. This temporal series can be regarded as the output of a Linear Time Invariant (LTI) system of unknown parameters [30].

It is well known [31] that, given a sequence of output measurements $\mathbf{A} = \{\mathbf{A}_i\}_{i=1}^{M}$, its associated truncated block-Hankel matrix is:

$$\widetilde{\mathbf{H}}_{\mathbf{A}} = \begin{bmatrix} \mathbf{A}_1, & \mathbf{A}_2, & \mathbf{A}_3, & \ldots, & \mathbf{A}_{m+1} \\ \mathbf{A}_2, & \mathbf{A}_3, & \mathbf{A}_4, & \ldots, & \mathbf{A}_{m+2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \mathbf{A}_{n-1}, & \mathbf{A}_n, & \mathbf{A}_{n+1}, & \ldots, & \mathbf{A}_M \end{bmatrix}, \quad (1)$$

where n is the maximal order of the system, M is the temporal length of the sequence, and it holds that $M = n + m - 1$. Finally, the Hankel matrix can be normalized as follows:

$$\mathbf{H}_{\mathbf{A}} = \frac{\widetilde{\mathbf{H}}_{\mathbf{A}}}{\sqrt{||\widetilde{\mathbf{H}}_{\mathbf{A}}\ \widetilde{\mathbf{H}}_{\mathbf{A}}^{T}||_F}}. \quad (2)$$

C. Creating Hankel Subspaces

In order to represent an ordered image set $\mathbf{A} = \{\mathbf{A}_i\}_{i=1}^{M}$ in terms of subspace and preserve spatial and temporal information, we introduce the concept of Hankel subspace for gesture recognition.

Subspace-based methods exploit the fact that a set of images lies in a cluster, which can be efficiently represented by a set of orthonormal basis vectors [15]. Our assumption is that the same formulation can be regarded for Hankel subspaces and, therefore we can achieve a novel representation for gesture-based image recognition.

Therefore, given a normalized Hankel matrix H_A from the ordered image set $\mathbf{A} = \{\mathbf{A}_i\}_{i=1}^{M}$, we can compute an autocorrelation Hankel matrix as:

$$\mathbf{C}_{\mathbf{A}} = \mathbf{H}_{\mathbf{A}}\mathbf{H}_{\mathbf{A}}^{T} \quad (3)$$

when $\mathbf{C}_{\mathbf{A}} \in \mathbb{R}^{k \times k}$, its eigendecomposition generates a set of eigenvectors $\mathbf{\Phi_A} = \{\phi_i\}_{i=1}^{k}$ that spans a subspace $\mathbf{P_A}$.

D. Selecting Samples

When creating a Hankel matrix, the number of images contained in a set and its dimension are crucial factors in terms of computational resources. In order to alleviate this issue, we introduce two approaches based on sample selection.

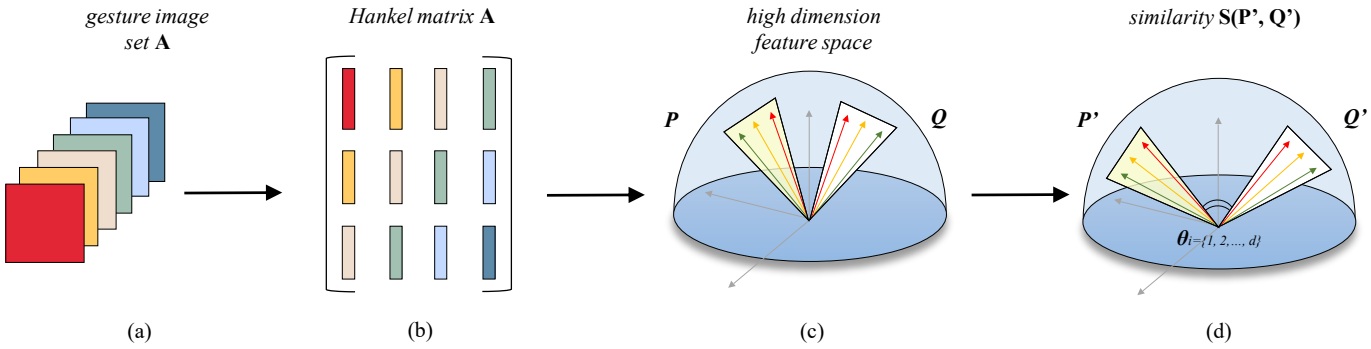

Fig. 1. Conceptual figure of our method. (a) An ordered subset of images representing a gesture $\mathbf{A}$ is handled, where a selection criterion is employed to reduce the number of images. (b) Then, the Hankel matrix $\mathbf{H_A}$ is created from the set of the selected images. (c) After that, we extract the basis vectors from the Hankel matrix $\mathbf{H_A}$ to produce the subspace $\mathbf{P}$ and its soft weights. Then, we orthogonalize the subspace to achieve a subspace $\mathbf{P}'$. (d) Finally, the soft weights are employed to achieve the structural similarity between $\mathbf{P}'$ and a reference subspace $\mathbf{Q}'$.

Random sample selection: In this approach, we randomly select images from the set, preserving its original order. We adopt this temporal sampling scheme in the image sequence since close images in time hardly change their appearance, containing high level of redundant information to identify the gesture that is being performed. This strategy also allows us to deal with sample reduction with a straightforward implementation.

Clustering selection: The second approach employs a clustering strategy, where the centroids obtained by $k-$means clustering are employed to represent the set, decreasing its number of images. The use of $k-$means clustering was previously employed for kernel dimensionality reduction in [32]. The advantage of using clustering is that the k centroids of the clusters will represent most of the relevant gesture information for discrimination, eliminating redundant images, achieving a good accuracy with low computational cost.

E. Computing the Soft Weights

In gMSM, all the eigenvectors are employed to represent a subspace. However, each eigenvector has its own weight, which is computed as follows; let $\mathbf{\Lambda}_m = \text{diag}(\lambda)$ be the eigenvalues of $\mathbf{C_A}$ in descending order, the design of the soft weights is performed according to these eigenvalues. Let $\mathbf{\Omega_A} = \text{diag}(w)$ be a diagonal matrix of soft weights:

$$\omega = w_m(\lambda) = \min\left[\frac{\lambda}{\lambda_m}, 1\right], \tag{4}$$

where w_m is the $m-$th eigenvalue in λ. This soft weighting evaluates the importance of each eigenvector as a basis in the subspace by the variance relative to λ_m. The m first values of the diagonal matrix $\mathbf{\Omega_A}$ will be unity and the remainder will be proportionally decreasing with the $m-$th eigenvalue.

In gMSM, each class subspace $\mathbf{P}_i$ uses the same parameter m. In general, this value is set from 1 to 4 in order to evaluate the importance of the eigenvectors in each subspace.

In contrast to gMSM, in HMS we employ an automatic approach to set the value of m. We adopt a heuristic based on the interpretation that eigenvectors corresponding to the eigenvalues larger than the average eigenvalues have high representative information. Let us denote λ_i as the $i-$th eigenvalue corresponding to the $i-$th eigenvector. The average eigenvalue μ_A is:

$$\mu_{\mathbf{A}} = \frac{1}{k}\sum_{i=1}^{k}\lambda_i. \tag{5}$$

Next, let us consider that λ_j is the smallest eigenvalue corresponding to the $j-$th eigenvector that satisfies $\lambda_j \leq \mu_{\mathbf{A}}$. Then, we set $m = j$. As in gMSM, these weights are unitary and the remainder eigenvectors will be proportionally decreased. This approach has several advantages. First, the computational cost required to set the m parameter is largely reduced, as we do not have to set m by parameter tuning. Second, each class subspace $\mathbf{P}_i$ will achieve a different set of weights $\mathbf{\Omega}_i$, regarding the spread of energy over the eigenvectors.

F. Orthogonalizing Hankel Subspaces

We will now explain the procedure to determine the orthogonalization matrix $\mathbf{W}$ in order to orthogonalize the c $m-$dimensional class subspaces with the orthogonal basis vectors $\{\mathbf{e}_i\}_{i=1}^{M}$ in the $n-$dimensional input space $\mathscr{I}$. This ortogonalization procedure enhances the difference between the Hankel subspaces, increasing the recognition rate of the framework.

Let the projection matrix corresponding to the projection onto the class i subspace $\mathbf{P}_i$,

$$\mathbf{P}_i = \sum_{i=1}^{M}\mathbf{e}_j\mathbf{e}_j^T, \tag{6}$$

where $\mathbf{e}_j$ is the $j-$th orthogonal basis vector of $\mathbf{P}_i$. Next, the total projection matrix is defined as:

$$\mathbf{G} = \sum_{i=1}^{r}\mathbf{P}_i. \tag{7}$$

By applying singular value decomposition on the total projection matrix $\mathbf{G}$, we obtain the $v \times n$ whitening matrix $\mathbf{W}$, defined by the following equation:

$$\mathbf{W} = \mathbf{\Lambda}^{-1/2}\ \mathbf{D}^T, \tag{8}$$

where v has dimension $r \times m$, (restricted to $v = n$, if $v > n$), $\mathbf{D}$ is the $n \times v$ matrix whose $i-$th column vector is the eigenvector of the matrix $\mathbf{G}$ corresponding to the $i-$th highest eigenvalue, and $\mathbf{\Lambda}$ is the $v \times v$ diagonal matrix with the $i-$th highest eigenvalue of the matrix $\mathbf{G}$ as the $i-$th diagonal component.

G. Hankel Subspaces Matching

After obtaining the Hankel subspaces and its weights, we can compute the similarity between the subspaces. This procedure is achieved by applying canonical angles or principal angles [33]. In subspace-based methods, we consider that if the distance between two subspaces is small enough, then we consider these subspaces similar to each other. Mathematically, let $\mathbf{\Phi_A} = \{\phi_i\}_{i=1}^k$ and $\mathbf{\Psi_Y} = \{\psi_i\}_{i=1}^k$ span two $k-$dimensional subspaces $\mathbf{P_A}$ and $\mathbf{Q_Y}$. Then, let $\mathbf{s}(\mathbf{P_A}, \mathbf{Q_Y}) = \{0 \leq \theta_1 \leq \theta_2 \leq \ldots \leq \theta_n \leq \pi/2\}$ represents the set of angles between $\mathbf{P_A}$ and $\mathbf{Q_Y}$.

A practical approach to determine $\mathbf{s}(\mathbf{P_A}, \mathbf{Q_Y})$ is by computing the $\mathbf{\Lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ eigenvalues of:

$$\mathbf{R} = \mathbf{\Omega_A}\ \mathbf{\Phi}_\mathbf{A}^T \mathbf{\Psi_B} \mathbf{\Omega_B}. \tag{9}$$

Then, the canonical angles:

$$\theta_i = \{\cos^{-1}(\lambda_1),\ \cos^{-1}(\lambda_2),\ \ldots,\ \cos^{-1}(\lambda_k)\}, \tag{10}$$

are employed to compute the structural similarity between soft weighted $\mathbf{\Phi_A}$ and $\mathit{\Psi}_\mathbf{B}$ Hankel subspaces as follows:

$$\mathbf{s}(\mathbf{\Phi_A}, \mathbf{\Psi_B})_M = \frac{1}{M} \sum_{i=1}^{M} \cos^2(\theta_i), \tag{11}$$

the structural similarities between Hankel subspaces are more robust to noise, such as illumination variations and point-of-view in sets of gesture images.

IV. Experimental Evaluation

In this section we show the experimental results of our proposed method. We employed Cambridge gesture dataset [34] for general gestures classification and Human-Computer Interaction (HCI) dataset [14], which contains computer interface gestures. In our experiments, we employed leave-one-out cross-validation. We report results for OHSM-I (random sample selection), OHSM-II ($k-$means sample selection) and gOHSM (generalized version of OHSM-II). We compare OHSM and variants with several state-of-the-art subspace-based methods: mutual subspace method (MSM) [15], discriminative canonical correlation analysis (DCC) [1], generalized mutual subspace method (gMSM) [18] and generalized difference subspace (GDS) [21]. As OHSM-I depends on a random selection and OHSM-II depends on the initial conditions of $k-$means clustering, for these methods, we performed each experiment 20 times. We report the average of these results.

TABLE I
Evaluated methods and its average accuracy.

Methods	Cambridge [34]	HCI [14]
MSM [15]	61.5%	56.7%
DCC [1]	82.0%	77.3%
gMSM [18]	75.5%	66.1%
GDS [21]	76.0%	71.4%
OHSM-I (random)	78.0%	74.8%
OHSM-II ($k-$means)	82.0%	77.5%
gOHSM	**85.5%**	**79.7%**

The Cambridge gesture dataset: consists of 9 classes of gestures. In total, there are 900 video sequences which are partitioned into 5 different illumination subsets. We reduce the size of the video frame to 20×20 pixels and then converted the images to grayscale. Each class contains 100 image sequences with 5 different illuminations and 10 arbitrary motions performed by 2 subjects.

Human-Computer Interaction (HCI) dataset: consists of both static and dynamic hand gestures according to mouse functionalities: cursor, left click, right click, mouse activation, and mouse deactivation. The dataset is divided into 2 sets, the first one has no information regarding the temporal segmentation of the frames and the second is properly segmented. In our experiments, we employed the second image set, where region of interest and label information are available. This set contains 30 labeled video sequences, which are performed by 6 different individuals, each video sequence contains in average 75 images. We reduce the size of the video frame to 20×20 pixels and then converted the images to grayscale.

Table I shows the results of the different evaluated methods for gesture recognition. Among the methods that do not employ Hankel matrices, DCC and GDS exhibit high discriminative power comparing to MSM and gMSM. This is a result that both DCC and GDS employ discriminative spaces, where more informative features may be extracted. On the other hand, MSM and gMSM rely only on affine subspaces, where no discriminative scheme is adopted.

In both datasets, we select $k = M/2$ images from each image set, achieving $k = 50$ images in Cambridge gesture dataset. In Human-Computer Interaction (HCI) dataset, the number of selected samples varies, as the image sets do not have the same number of images. In average, $k = 37$ images. For the random selection schema, we use the same number of images as employed for the clustering sample selection.

OHSM-I achieved competitive accuracy, similar to DCC. This indicates that the temporal information extracted by the Hankel representation is very powerful, even when random samples are selected, the main concern here is that the selected samples should preserve its temporal order. OHSM-II achieved a higher accuracy than OHSM-I, demonstrating that $k-$means clustering is more efficient than random sample selection. This is an expected result, as selecting the centroids obtained by $k-$means is more likely to preserve the structural

Fig. 2. On the left: ample images from the Cambridge Hand Gesture dataset. On the right: Sample images from the Human-Computer Interaction dataset.

information of the gesture manifold than random selection. gOHSM achieved the highest accuracy among the evaluated methods, indicating that the weighted structural similarity between subspaces extracted from Hankel matrices is very efficient for gesture recognition from image sets.

From the Table I we observe that all the methods presented a sharp drop in accuracy when comparing the results from the Cambridge dataset and HCI dataset. This is a consequence of the different background from each dataset. In Cambridge dataset, the gesture images where collected in a controlled background, different from the HCI dataset, where the images were recorded in an unconstrained background.

As final remark, we would like to emphasize that OHSM and variants do not employ any learning scheme, different from DCC and GDS, where a discriminant space is employed in order enhance the discriminability among the gesture classes. This demonstrates the effectiveness of employing Hankel subspace for gesture representation.

V. Conclusions

We presented a novel framework for gesture recognition from sets of images. In our work, we introduced a compact representation based on sample selection, Hankel matrix and automatic soft weighting. In order to reduce the number of samples per image set class, we evaluate two strategies. In the first one, we randomly select the image samples. In the second one, we select the most representative samples by using k-means clustering. The selected patterns are then adopted to construct a Hankel matrix in order to preserve the temporal relation between the image patterns. By extracting the eigenvectors from the Hankel matrix, we achieved a high compact representation. In order to efficiently use all of the basis vectors from the Hankel matrix, we proposed a modified version of the soft weighs. Instead of manually select the number of eigenvalues, where its weights are set to 1, we show by experiments that using the average of the eigenvalues as a soft threshold we accomplish high recognition rates, without searching for all the basis vector combinations. This novel approach shows higher recognition rate compared to the conventional method. Finally, comparing the sample selection strategies, the random selection demonstrated high efficient time and competitive recognition rate. By selecting the samples using $k-$means clustering, we achieved superior recognition rate compared to all of the investigated methods.

For future directions, we will extend our work to deal with nonlinear patterns, as the subspaces employed in this work are mainly based on linear transformations. One can achieve this objective by using a nonlinear variant of PCA, as kernel PCA. Another research avenue would use a discriminative approach in order to optimally select the samples from each class. In our method, we select the samples essentially by its representative importance in terms of class distribution, without taking into consideration the discriminative power in relation to the other classes. We understand that by using a discriminative selection criterion one can improve the recognition rate of the framework.

References

[1] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, 2007.

[2] T.-K. Kim, S.-F. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[3] Z. Yang, Y. Li, W. Chen, and Y. Zheng, "Dynamic hand gesture recognition using hidden markov models," in *Proc. Int. Conf. on Computer Science & Education (ICCSE)*, 2012, pp. 360–365.

[4] Y. Yin and R. Davis, "Real-time continuous gesture recognition for natural human-computer interaction," in *Proc. Int. Symp. on Visual Languages and Human-Centric Computing (VL/HCC)*, 2014, pp. 113–120.

[5] H.-S. Yeo, B.-G. Lee, and H. Lim, "Hand tracking and gesture recognition system for human-computer interaction using low-cost hardware," *Multimedia Tools and Applications*, vol. 74, no. 8, pp. 2687–2715, 2015.

[6] A. Hernández-Vela, M. Á. Bautista, X. Perez-Sala, V. Ponce-López, S. Escalera, X. Baró, O. Pujol, and C. Angulo, "Probability-based dynamic time warping and bag-of-visual-and-depth-words for human gesture recognition in rgb-d," *Pattern Recognition Letters*, vol. 50, pp. 112–121, 2014.

[7] A. Birdal and R. Hassanpour, "Region based hand gesture recognition," 2008.

[8] M. R. Malgireddy, I. Inwogu, and V. Govindaraju, "A temporal bayesian model for classifying, detecting and localizing activities in video sequences," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012, pp. 43–48.

[9] P. Pławiak, T. Sośnicki, M. Niedźwiecki, Z. Tabor, and K. Rzecki, "Hand body language gesture recognition based on signals from specialized glove and machine learning algorithms," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 1104–1113, 2016.

[10] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the accuracy and robustness of the leap motion controller," *Sensors*, vol. 13, no. 5, pp. 6380–6393, 2013.

[11] J. Konecnỳ and M. Hagara, "One-shot-learning gesture recognition using hog-hof," *Journal of Machine Learning Research*, vol. 15, pp. 2513–2532, 2014.

[12] P. Sykora, P. Kamencay, and R. Hudec, "Comparison of sift and surf methods for use on hand gesture recognition based on depth map," *AASRI Procedia*, vol. 9, pp. 19–24, 2014.

[13] J. Wan, G. Guo, and S. Z. Li, "Explore efficient local features from RGB-d data for one-shot learning gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1626–1639, 2016.

[14] A. I. Maqueda, C. R. del Blanco, F. Jaureguizar, and N. García, "Human–computer interaction based on visual hand-gesture recognition using volumetric spatiograms of local binary patterns," *Computer Vision and Image Understanding*, vol. 141, pp. 126–137, 2015.

[15] K.-i. Maeda, "From the subspace methods to the mutual subspace method," in *Computer Vision*. Springer, 2010, pp. 135–156.

[16] B. B. Gatto, L. S. de Souza, and E. M. dos Santos, "A deep network model based on subspaces: A novel approach for image classification," in *Machine Vision Applications (MVA), 2017 Fifteenth IAPR International Conference on*. IEEE, 2017, pp. 436–439.

[17] J. Lu, Y.-P. Tan, and G. Wang, "Discriminative multimanifold analysis for face recognition from a single training sample per person," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 39–51, 2013.

[18] T. Kobayashi, "Generalized mutual subspace based methods for image set classification," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 578–592.

[19] B. B. Gatto, S. Waldir, M. Eulanda, and D. Santos, "Kernel two dimensional subspace for image set classification," in *Proc. Int. Conf. on Tools with Artificial Intelligence (ICTAI)*, 2016, pp. 1004–1011.

[20] B. B. Gatto and E. M. dos Santos, "Image-set matching by two dimensional generalized mutual subspace method," in *Proc. Brazilian Conf. on Intelligent Systems (BRACIS)*. IEEE, 2016, pp. 133–138.

[21] K. Fukui and A. Maki, "Difference subspace and its generalization for subspace-based methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 11, pp. 2164–2177, 2015.

[22] F. Chatelin, *Eigenvalues of Matrices: Revised Edition*. SIAM, 2012.

[23] H. Itoh, A. Imiya, and T. Sakai, "Dimension reduction and construction of feature space for image pattern recognition," *Journal of Mathematical Imaging and Vision*, vol. 56, no. 1, pp. 1–31, 2016.

[24] R. Zhu, K. Fukui, and J.-H. Xue, "Building a discriminatively ordered subspace on the generating matrix to classify high-dimensional spectral data," *Information Sciences*, vol. 382, pp. 1–14, 2017.

[25] B. Li, O. I. Camps, and M. Sznaier, "Cross-view activity recognition using hankelets," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1362–1369.

[26] L. Lo Presti and M. La Cascia, "Using hankel matrices for dynamics-based facial emotion recognition and pain detection," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 26–33.

[27] Y. Lin, K. Abdelfatah, Y. Zhou, X. Fan, H. Yu, H. Qian, and S. Wang, "Co-interest person detection from multiple wearable camera videos," in *Proc. Int. Conf. on Computer Vision*, 2015, pp. 4426–4434.

[28] K. Sugimoto and S.-i. Kamata, "Fast color matching using weighted subspace on medicine package recognition," in *Proc. Int. Conf. on Machine Vision Applications (MVA)*, 2011, pp. 287–290.

[29] A. Mian, Y. Hu, R. Hartley, and R. Owens, "Image set based face recognition using self-regularized non-negative coding and adaptive distance metric learning," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 5252–5262, 2013.

[30] E. Sontag, "Nonlinear regulation: The piecewise linear approach," *IEEE Transactions on Automatic Control*, vol. 26, no. 2, pp. 346–358, 1981.

[31] M. Viberg, "Subspace-based methods for the identification of linear time-invariant systems," *Automatica*, vol. 31, no. 12, pp. 1835–1851, 1995.

[32] K. Hotta, "Local co-occurrence features in subspace obtained by kpca of local blob visual words for scene classification," *Pattern Recognition*, vol. 45, no. 10, pp. 3687–3694, 2012.

[33] F. Chatelin, *Eigenvalues of Matrices: Revised Edition*. SIAM, 2012, vol. 71.

[34] T.-K. Kim and R. Cipolla, "Canonical correlation analysis of video volume tensors for action categorization and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1415–1428, 2009.