Covariance Profiles: A Signature Representation For Object Sets

Anoop K.R.*, Adway Mitra†, Ujwal Bonde*, Chiranjib Bhattacharyya†, K.R.Ramakrishnan*
*Electrical Engineering, †Computer Science and Automation, IISc, Bengaluru
{{anoopkr,krr}@ee,{adway,chiru}@csa}.iisc.ernet.in

Abstract

*We consider the problem of extracting a signature representation of similar entities employing covariance descriptors. Covariance descriptors can efficiently represent objects and are robust to scale and pose changes. We posit that covariance descriptors corresponding to similar objects share a common geometrical structure which can be extracted through joint diagonalization. We term this diagonalizing matrix as the **Covariance Profile** (CP). CP can be used to measure the distance of a novel object to an object set through the **diagonality measure**. We demonstrate how CP can be employed on images as well as for videos, for applications such as face recognition and object-track clustering.*

1 Introduction

With the advent of sites such as *YouTube*, *Picasa* and *Flickr*, there has been an explosion of visual content on the internet. However, this has also resulted in enormous redundancy of information owing to content duplication. Efficient representation of visual content can minimize storage requirements and enable efficient comparison of images and videos, facilitating tasks like face recognition, video clustering and retrieval.

Efficient object representation has been well studied in computer vision. Many object comparison techniques fail in the presence of pose, scale or illumination changes. In this respect, covariance descriptors [8] have been generally found to be stable, robust to pose or scale variations and also provide for efficiently fusing multiple features by capturing between-feature relationships.

Principal Angles (PA) [11] are popularly employed for comparing object sets. The underlying assumption in PA is that instances belonging to the same entity span a linear subspace. Two sets are compared by measuring the angle between their subspaces. Kernelized principal angles (KPA) has also been discussed in [9]. However, they don't allow for compact data representation and the subspace constraint makes them suitable only for matching sets (as against individual elements).

We propose the **covariance profile** (CP), a novel *signature descriptor* that compactly represents a set of similar objects. The intuition behind CP is that the same principal directions are shared by similar objects. These directions are obtained by simultaneously diagonalizing the covariance matrices corresponding to the individual objects. This paper represents the first work employing CP as a concise *object-set representation*. We demonstrate how CPs are useful for a) object-track clustering (where a *object-track* denotes regions corresponding to an object tracked in a video) and b) image-based face recognition on badly aligned and cropped faces.

2 Covariance Profiles

This section formally defines a covariance profile and discusses its extraction.

2.1 Definition and Estimation of Covariance Profile

Consider a set of similar objects $\{I_1, I_2, \ldots, I_N\}$, described by their Covariance Descriptors as set $T = \{C_1, C_2, \ldots, C_N\}$ henceforth referred as a *family*. We attempt to capture the similarity structure for the family with a set of vectors $\beta_1, \beta_2, \ldots, \beta_d$, where d is the number of columns of the C_i's, such that

$$C_i = \sum_j \lambda_{ij} \beta_j \beta_j^T \tag{1}$$

We consider the matrix V, with vectors $\{\beta_j\}$ as its columns. V *Jointly Diagonalizes* the individual C_i matrices, i.e. $\Lambda_i = V^T C_i V$ is a diagonal matrix whose diagonal entries are λ_{ij}. In practice, such a V may not exist. However, it is possible to compute a V which will *approximately diagonalize* the $\{C_i\}$ matrices. This matrix is defined as the **covariance profile** for the family.

To estimate a CP of a given family, we make use of *approximate joint diagonalization* algorithms. Different formulations have been reviewed and discussed in [12]. We estimate CP using Pham's algorithm [7] which is designed for joint diagonalization of positive definite Hermitian matrices. At each step, the algorithm

proceeds by performing successive transformations on rows l, m of V, according to

$$\begin{bmatrix} V_l \\ V_m \end{bmatrix} = F_{lm} \begin{bmatrix} V_l \\ V_m \end{bmatrix} \quad (2)$$

where F_{lm} is a 2×2 non-singular matrix such that (3) is sufficiently reduced. Here V need not be orthogonal.

$$\sum_i score(V, C_i) \quad (3)$$

$$\text{where } score(V, C_i) = log(\frac{det(diag(V^T C_i V))}{det(V^T C_i V)}) \quad (4)$$

Algorithm consists of repeated *sweeps* till convergence.

2.2 Distance Measures based on CP

The *diagonality score* is a measure of diagonality of the matrix C, with respect to the CP, V. It is shown in [3] that this value decreases with decreasing Frobenius norm of the off-diagonal elements in $V^T C V$, and is 0 if and only if it is fully diagonal. Consider a family T with CP V. Given a new covariance matrix C, its closeness to the family T can be computed using the diagonality measure (4). A sufficiently small diagonality score indicates that C is almost perfectly diagonalized by V, and hence likely to belong to the family T. We also compare a family to a CP V as

$$score(V, T) = \min_i score(V, C_i) \quad (5)$$

If V_i and V_j denote the CPs of $T_i = \{C_{i1}, C_{i2}, \ldots, C_{iN_i}\}$ and $T_j = \{C_{j1}, C_{j2}, \ldots, C_{jN_j}\}$ respectively then the distance between the families is measured as

$$score(T_i, T_j) = \min(score(V_i, T_j), score(V_j, T_i)) \quad (6)$$

3 Applications

In this section, to evaluate CP we demonstrate how they can be used for clustering object-tracks in videos and for giving a signature representation to a person as in face recognition.

3.1 Object-track Clustering

A video consists of successive frames captured over a period of time with temporally adjacent frames being similar. We define an *object-track* (or simply, a *track*) as a set of images obtained from successive video frames, such that each of these images contain a unique object-these images can be the cropped outputs of an object detector or a tracker. Frames from a car video and the *track* corresponding to the car are shown in Fig 1. **Object-track clustering** is the task of grouping *tracks* such that all *tracks* assigned to a particular cluster correspond to the same entity.

Let each frame of the *track* be represented by R overlapping region covariance descriptors. We denote T_i^r as that part of the i^{th}*track* corresponding to region r, *i.e.*

$$T_i^r = \{C_{il}^r\} \; \forall \; l = 1, 2, ..., n_i, \; r = 1, 2, ...R \quad (7)$$

Here, n_i is the number of frames in track T_i while l denotes frame number. A total of R regions are considered for representing each frame of the *track*. In our experiments, we fix $R = 5$ and the corresponding regions for a face are as shown in Fig 2.

Figure 1. Top: A car tracked in consecutive frames; Bottom: Regions in bounding box of the track forms the object-track

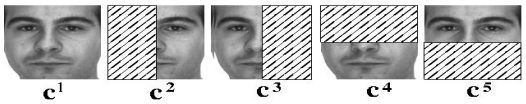

Figure 2. Five face regions considered .

Let the CP associated with track T_i^r be V_i^r . Distance measure between two tracks T_i and T_j is defined by

$$d(T_i, T_j) = \sum_{r=1}^{R} score(T_i^r, T_j^r) - \max_r score(T_i^r, T_j^r) \quad (8)$$

This is converted to a similarity measure as

$$S(T_i, T_j) = e^{-d(T_i, T_j)/2b} \quad (9)$$

where b is a constant.

To evaluate the clustering, we use the **purity** measure. Intuitively, we want each group to be **pure**, *i.e.*, all the tracks assigned to the group should be associated with the same entity, even though a particular entity may generate more than one cluster. Let the total number of clusters be m. To a particular cluster k, we assign an entity label L_k as

$$L_k = \arg\max_e \sum_{1 \le i \le N_k} \delta_{iek} \quad (10)$$

where $\delta_{iek} = 1$ if the i^{th} track in cluster k belongs to entity e, and N_k is the number of *tracks* in k^{th} cluster. In other words, we find the most frequently assigned entity label for all tracks in the cluster. Next, we define the clustering purity for m clusters as

$$P(m) = \frac{1}{N} \sum_{1 \le k \le m} \sum_{1 \le i \le N_k} \delta_{iL_k k} \qquad (11)$$

where N is the total number of tracks. Finally, if experiments are done over different cluster configurations $\{m_1, m_2, \ldots, m_k, \ldots, m_K\}$, the maximum clustering purity is defined as $MC = \max_k P(m_k)$

3.1.1 Experiments and Results

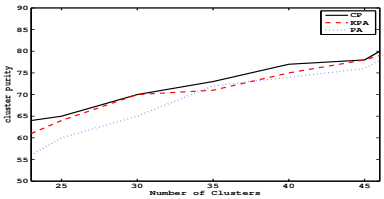

Figure 3. Plot of cluster purity v/s Number of clusters for YouTube celebrity dataset

Dataset	CP	PA	KPA
Sitcom	90.57(10)	88.86(10)	90.76(10)
YouTube	80.35(46)	78.60(46)	79.7 (46)
Objects	69.62(15)	64.62(16)	65 (15)

Table 1. Comparison of maximum clustering purity for different methods. Numbers in brackets denote number of clusters for which clustering purity is maximum.

Datasets: We used 3 datasets for our experiments. The first dataset consists of 175 clips corresponding to 5 actors (35 clips per actor) obtained from a television sitcom. The length of each clip is between 10-350 frames. The second consists of 789 *YouTube* clips taken from [4] corresponding to 23 celebrities containing their respective face. However, the face pose can vary from frontal to profile. The third comprises 260 short *YouTube* videos corresponding to 8 different objects.

Each of these clips is associated with a single entity (*person* for first two datasets and *object* for the third). We created tracks from each clip by detecting/tracking this entity [1, 2]. 40-dimensional Gabor features (5 scales and 8 orientations) were used to compute covariance features for each region as in [8].

After obtaining tracks for all clips we calculate the similarity between each pair using (9). Thus, we obtain a symmetric matrix S, where $S(i, j)$ specifies the similarity between tracks i, j. Spectral clustering [5] is then performed on S. For K entities, we vary the number of clusters from K to $2K$. We compare our proposed similarity measure using CP against principal angles [11] and Kernelized Principal Angles (KPA) [9] with spectral clustering. Comparable results with respect to computationally intensive KPA are obtained for the first two datasets, while we outperform both PA and KPA-based approaches for the 'Objects' dataset. The results are as shown in Table 1. Fig 3 shows the variation in purity-with the number of clusters. In general, purity improves with number of clusters because multiple, tighter clusters are generated corresponding to each entity.

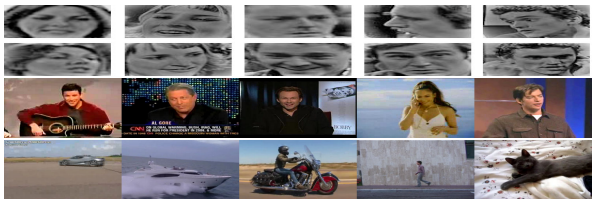

Figure 4. Sample frames from the first (rows 1 and 2), second (3rd row) and third (bottom-row) datasets.

3.2 Face Recognition

In this subsection, we demonstrate how CPs are useful for discriminating between different instances of the same entity. We consider the problem of image-based face recognition where the images are *badly aligned and cropped*.

3.2.1 Procedure

Each face is represented by R covariance matrices, as described earlier. Thus, from the training set for a class we obtain R CPs representing the class. Classification is performed using the diagonality scores, eqn. (4), of the covariance matrices corresponding to a test face with the CPs of each class as illustrated in Algorithm 1.

3.2.2 Experiments and Results

Our approach can efficiently perform recognition with badly aligned and cropped faces, where traditional approaches fare poorly. We compare our proposed distance measure using CP (4) against two other methods which can also work under similar conditions. The first approach by Wright *et al.* [10] employs sparse modeling, which is efficient independent of the features used. The second approach employs Gabor-based covariance features [6] for classification based on geodesic distance.

We use standard AR and YaleB databases. The faces used in our experiment are cropped using a face detector and are not preprocessed. The AR database contains

Algorithm 1 CLASSIFICATION ALGORITHM

TRAINING ALGORITHM

Number of Classes= N

for each $i = 1 : N$

Represent each training image in class i with R Covariance Matrices

for each $r = 1 : R$

Obtain and save CPs V_i^r

end for

end for

TESTING ALGORITHM

Represent the test image I by R covariance matrices C_I^r where $r = 1 : R$.

for $i = 1 : N$

Calculate $d_{iI}^r = score(V_i^r, C_I^r), \forall\, r = 1 : R$.

end for

$Class(I) = \underset{i=1 \rightarrow N}{\arg\min}(\sum_{r=1}^{R} d_{iI}^r - \max_r\ d_{iI}^r)$

faces of 110 subjects. For each subject, we consider 14 faces without occlusion, of which, seven are respectively used for training and testing. The YaleB database consists of 28 subjects, and 10 randomly chosen images are used for training and testing without any overlap between the two sets.

The recognition results are presented in Figure 5. For the sparse modeling approach, results are shown for the dimension that produces the maximum accuracy. Evidently, our approach outperforms competing methods for poorly cropped and aligned faces, demonstrating the robustness of covariance profiles. In Figure 6 our approach shows superior recognition performance with increasing training samples for the YaleB dataset.

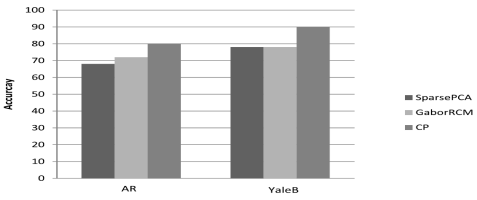

Figure 5. Plot of face recognition accuracy for AR and YaleB database.

4 Conclusions

This paper introduces covariance profiles (CPs), which is a novel signature representation for an object-set. The object-set, or *family*, can contain many instances of the same entity/similar entities. CPs provide for an efficient comparison of novel objects with the family through the diagonality score. The utility of CPs

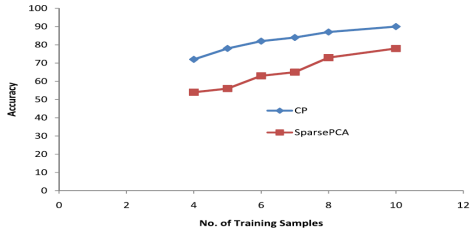

Figure 6. Accuracy with increasing training samples for YaleB dataset

is demonstrated for object-track clustering in videos, as well as for image-based face recognition. We observe that the performance obtained employing CPs is comparable to/superior than competing methods for both the test scenarios.

References

[1] K. R. Anoop, P. Anandathirtha, K. R. Ramakrishnan, and M. S. Kankanhalli. Integrated detect-track framework for multi-view face detection in video. *ICVGIP*, pages 336–343, 2008.

[2] D.Comaniciu, V.Ramesh, and P.Meer. Kernel-based object tracking. *PAMI*, 25(5):564–577, 2003.

[3] B. N. Flury. Common principal components and related multivariate models. *John Wiley and Sons*, 1988.

[4] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. *CVPR*, pages 1–8, 2008.

[5] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *NIPS*, 14:849–856, 2001.

[6] Y. Pang, Y. Yuan, and X. Li. Gabor-based region covariance matrices for face recognition. *CSVT*, 18(7):989–993, 2008.

[7] D. T. Pham. Joint approximate diagonalization of positive definite hermitian matrices. *SIAM J. Matrix Anal. Appl.*, 22(4), 2001.

[8] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. *ECCV*, pages 589–600, 2006.

[9] L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *JMLR*, 4:931, 2003.

[10] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *PAMI*, 31(2):210–227, 2009.

[11] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. *FGR*, pages 318–323, 1998.

[12] A. Ziehe, P. Laskov, G. Nolte, and K. Muller. A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. *JMLR*, 5:777–800, 2004.