Spatial Interest Pixels (SIPs): Useful Low-Level Features of Visual Media Data

Qi Li
Department of CIS
University of Delaware
qili@cis.udel.edu

Jieping Ye
Department of CS
University of Minnesota
jieping@cs.umn.edu

Chandra Kambhamettu
Department of CIS
University of Delaware
chandra@cis.udel.edu

Abstract

Visual media data such as an image is the raw data representation for many important applications. The biggest challenge in using visual media data comes from the extremely high dimensionality. We present a comparative study on spatial interest pixels (SIPs), including eight-way (a novel SIP miner), Harris, and Lucas-Kanade, whose extraction is considered as an important step in reducing the dimensionality of visual media data. With extensive case studies, we have shown the usefulness of SIPs as the low-level features of visual media data. A class-preserving dimension reduction algorithm (using GSVD) is applied to further reduce the dimension of feature vectors based on SIPs. The experiments showed its superiority over PCA.

1 Introduction

Visual media data such as an image is the raw data representation for many important applications, such as facial expression recognition [18, 29], face recognition [27, 20, 24], video classification [13], etc. The biggest challenge in using the visual media data comes from its extremely high dimensionality.

To reduce the dimensionality of visual media data, the first step is usually to extract the *low-level features*[1]. Color, texture, shape/contour are three types of low-level features frequently used [26, 3, 25, 7]. The use of interest pixels [2] (say corners, salient image points) has attracted attention [23, 15] because of their *repeatability* (an interest pixel found in one image can be found again in another if these two images are spatially similar to each other). Interested readers can further refer to [16, 12].

We study spatial interest pixels (SIPs) in this paper. Intuitively, a SIP is a pixel that has stronger interest strength than most of other pixels in an image. The interest strength is basically measured by the change in pixel values along different 2D directions (say horizontal, vertical, etc). We study three miners of SIPs, eight-way, Harris, and Lucas-Kanade. The latter two are commonly used in computer vision community [9, 17] (Harris is attracting more attention in image retrieval community [23]). Both Harris and Lucas-Kanade utilize the local change of pixel values along only two directions (left-to-right and top-to-bottom), and non-maximum suppression is applied to suppress the pixels that do not have the strongest interest strength within their neighborhood. Eight-way is a novel SIP miner that utilizes the local change of pixel values along eight directions (uniformly distributed from 0 to 360 degrees). The distributions of SIPs (over the regular grid of image plane) is then used as feature vectors for classification task.

The dimensionality of SIP distributions is still pretty high (it ranges from several hundreds to several thousands). A class-preserving dimension reduction algorithm is thus introduced to further reduce their dimensionality. The class-preserving dimension reduction algorithm is based on the generalized discriminant analysis. The difference between generalized discriminant analysis and classic linear discriminant analysis (LDA), also called Fisher discriminant analysis (FDA) [6], is the use of the trace optimization in the former. GSVD provides a convenient tool for generalized discriminant analysis.

To show the usefulness of SIPs as low-level features for visual media data, we present our results on universal facial expression recognition [5] and face recognition. We tested one facial expression dataset Jaffe[3][18], four face datasets, including Jaffe, Yale[4][2], AR1[5][19], and Stirling[6][8]. Our results are very encouraging (see Section 7). The classifier used in this paper is nearest neighbor [1].

[1] We will strictly distinct the term feature from the term feature vector in the context of media-based classification applications. The former means color, texture, shape and pixel, whereas the latter means the representation of an image/video instance that are ready to feed into some classifier.

[2] In the context of image retrieval or 3D computer vision, they are called *interest points*. Renaming them as interest pixels in the context of data mining is to avoid confusion between the image point and data point (i.e., feature vector).

[3] http://www.mis.atr.co.jp/~mlyons/jaffe.html

[4] http://cvc.yale.edu/projects/yalefaces/yalefaces.html

[5] http://rvl1.ecn.purdue.edu/~aleix/aleix_face_DB.html

[6] http://pics.psych.stir.ac.uk/

Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03)
0-7695-1978-4/03 $ 17.00 © 2003 IEEE

The rest of the paper is organized as follows: Section 2 presents the related work. Section 3 reviews Harris and Lucas-Kanade detectors, and presents a novel SIP miner. Section 4 is on SIP distributions. A class-preserving dimension reduction algorithm is presented in Section 5. Nearest neighbor as the unique classifier is briefly reviewed in Section 6. Section 7 presents three case studies on SIP feature vectors. Finally, we conclude our study on SIP and give a future work in Section 8.

2 Related work

[23] gave a survey on several interest pixel detectors. To evaluate the performance of an interest pixel detector, two criteria are used. One is the repeatability, and the other is entropy. The criterion of repeatability is important in our work, mostly for the usefulness of SIP distribution. We did not give quantitative measure of the repeatability of SIP in this paper (but the classification accuracy in case studies gives a qualitative measure of repeatability). Interested readers can find more details in [23]. [15] presented a wavelet based salient pixel detector. Their motivation comes from the multiresolution property of wavelet coefficients.

Principal component analysis (PCA) is a widely used dimension reduction [11]. PCA does not utilize the class label information of training data. The limitation of PCA can be overcome by introducing linear/Fisher discriminant analysis (LDA/FDA) [6] where within-class scatter and between-class scatter are used to refine the single global covariance matrix in PCA. The classic LDA involves the inverse computation of one scatter matrix. If that matrix is singular, classic LDA will fail. We consider the generalized discriminant analysis whose criterion is to minimize the ratio of the traces of within-class scatter to between-class scatter. Using generalized singular value decomposition (GSVD), we can derive a simple algorithm to solve the trace ratio minimization problem and thus achieve class-preserving dimension reduction. [22] have the similar work on GSVD-based discriminant analysis, but they did not explicitly define any global objective function.

Static universal facial expression recognition is still a hard problem. [29] obtained the accuracy of around 90% using 10-fold cross validation [7], Gabor wavelet coefficients at 34 manually extracted fiducial points as the feature vectors, neural network as the classifier, and Jaffe as the test dataset. With the same Gabor feature and same evaluation method used in [29], [18] applied LDA to Jaffe dataset and obtained the accuracy of 92%. The disadvantage of their method is that the feature vectors are essentially manually extracted.

Intensive research on face recognition has been done in the past 30 years [27, 4, 20, 24]. It is beyond our ability to give an even relatively comprehensive view on the problem. Two surveys on face recognition [4, 30] can be excellent sources for the interested readers. The most recent work on the comparison between PCA and LDA can be found in [20] that used pixel values of face regions as feature vectors, nearest neighbor (in L_2-norm) as classifier, AR as the test dataset (AR is further split into several datasets). With different further split datasets, the accuracies range from 60% to 90%.

3 SIP miners

In this section, we will first review the Harris and Lucas-Kanade detectors, including the basic concept, algorithm implementation, and parameter setting. Based on another natural perspective in interpreting a spatial interest pixel, we then present the eight-way miner.

3.1 Harris and Lucas-Kanade

Given an image I and a pixel p, assume $I_x(p)$ and $I_y(p)$ are the derivatives at p along x and y axis. Both Harris detector and Lucas-Kanade detector are based on the gradient correction matrix of p defined as the following formula:

$$C(p) = \begin{pmatrix} \sum\limits_{q \in O_p} w_q I_x^2(q) & \sum\limits_{q \in O_p} w_q I_x(q) I_y(q) \\ \sum\limits_{q \in O_p} w_q I_x(q) I_y(q) & \sum\limits_{q \in O_p} w_q I_y^2(q) \end{pmatrix}, \tag{3.1}$$

where O_p is a square neighborhood of p, and $(w_q)_{q \in O_p}$ is a 2D-smoothing filter used to weight the derivatives.

In Harris detector, the interest strength of pixel p is defined to be the summation of the eigenvalues of $C(p)$. To save the computational time, the following equivalent form is more commonly used in practice [23]:

$$\text{strengh}(p) = \det(C(p)) - \alpha \text{trace}^2(C(p)), \tag{3.2}$$

where α is a discriminant factor that is usually set to be 0.6 [23]. In Lucas-Kanade detector, the interest strength of p is defined to be the minimal eigenvalue of $C(p)$. After the interest strength of each pixel in an image is computed, all pixels will be sorted according to their strength, and the h pixels with highest interest strength will be chosen as the SIPs. To spread out the SIPs, *non-maximum suppression* is applied. More specifically, for each pixel p, we compare its interest strength and the interest strength of its neighboring pixels (usually defined to be those pixels in a small square window (say 3 × 3) centered by p; if it is not the maximum, then its strength is reset to be zero (i.e., be suppressed).

Note that using the gradient correlation matrix of the derivatives rather than the derivatives themselves to decide the interest strength of a pixel is to earn the invariance to image orientation.

[7] In 10-fold cross validation, an entire dataset will first be split into 10 pieces. Then the test will be run 10 times. In each time, 9 pieces are used as training data, and remaining one piece is used as test data. The final accuracy estimation is the mean estimation.

Even though the algorithm implementation of Harris detector and Lucas-Kanade detector is simple, their parameter setting plays more important roles. There are two versions on the parameter setting of Harris detector (more details can be found in [23]). In the standard Harris, the derivative I_x or I_y is computed by convolution with the mask [-2,-1,0,1,2], and the filter used in weighting the derivatives is a Gaussian ($\sigma = 2$). In an improved version of Harris [23], I_x or I_y are computed by replacing the [-2,-1,0,1,2] mask by derivatives of a Gaussian ($\sigma = 1$). The improved version of Harris detector is found to mine the interest pixels with highest repeatability in the comparative study in [23] In our case studies, we will use the improved version of Harris. We now summarize the Harris and Lucas-Kanade detectors as Algorithm 1:

Algorithm 1 Harris/Lucas-Kanade SIP miner

```
Input: An image
Output: SIPs
  1. Compute the image gradient
  2. For each pixel $p$
     2.1. Form matrix $C(p)$ by formula (3.1)
     2.2. (Harris) Strength is assigned according to formula (3.2)
     $2.2'$. (Lucas-Kanade) Strength is assigned by the smaller eigenvalue of $C(p)$
  3. Non-maximum suppression
  4. Choose the first $n$ pixels of largest strength
```

3.2 Eight-way

Intuitively, if a pixel is spatially interesting, its rich information may not be sufficiently covered by the derivatives along two directions. Fig. 1 demos three typical and simplified cases that a SIP may look like. The changes of pixel values significantly across the edge/boundary, but changes slightly within a same region. A long (short) arrow shows the large (small) derivative magnitude along the direction of the arrow. The common ground of these three cases is that these pixels (expected to be SIPs) have a majority of long arrows (i.e., the numbers of long arrows are always more than 4). This observation gives us a new interpretation of SIP: *if the number of strong derivative is larger than 4, then the pixel is a good candidate of SIP and will be assigned a derivative-related value as interest strength; otherwise it will be assigned zero strength.* To distinguish between a large derivative and a small derivative, we first compute the mean of eight derivatives and then claim those above the mean to be large ones. The relatively challenging part in the above definition of SIP is the assignment of interest strength. Apparently, we at least have two options: one is maximal derivative, and the other is the mean derivative. We currently choose the first option based on the visual evaluation of the distribution of SIPs in several experiments of face images.

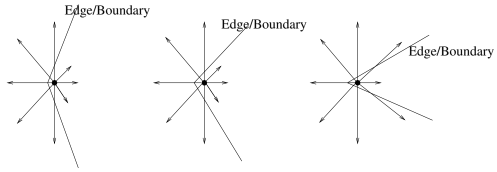

Figure 1. Three typical cases showing the asymmetry of SIPs

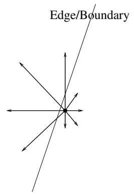

Figure 2. The symmetry of pixels near an edge/boundary.

Fig. 2 shows that the eight-way SIP miner has strong ability in discriminating an edgel from a SIP because an edgel has equal number of long arrows and short arrows. So non-maximum suppression is not necessary for eight-way miner. It is good for SIP mining because non-maximum suppression involves local uncertainty. Algorithm 2 summarizes eight-way SIP miner.

Algorithm 2 Eight-way SIP miner

```
Input: An image
Output: SIPs
  1. For each pixel $p$
     1.1. For each of eight directions
          1.1.1. Compute the change of along that direction (by convolving
          a Gaussian first derivative)
     1.2. Compute the mean change
     1.3. Count the number of changes above mean change
     1.4. If the count is larger than 4
          strength is assigned to be the largest change
          otherwise
          strength is set to 0
  2. Choose the first $h$ pixels of largest strength
```

3.3 SIP Examples

Fig. 3 shows the 300 SIPs found by applying the three SIP miners, eight-way miner, improved version of Harris detector (for convenience, we will still call it Harris in the rest of the paper), and Lucas-Kanade detector, on two different face images (namely, KA and KL that are taken from Jaffe dataset).

Let us first consider the SIPs on KA (in the first row).

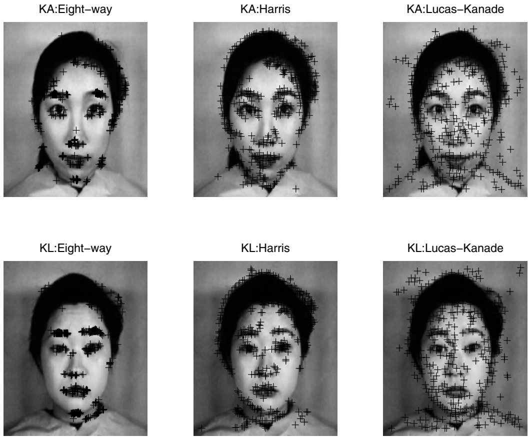

Figure 3. SIPs found by applying eight-way, Harris, and Lucas-Kanade detector on two different faces, namely KA and KL.

The SIPs found by eight-way are mostly located near the facial features (say eye, mouth, etc) of KA's face. Among the SIPs found by Harris, there are many located along some edges/boundaries (of large brightness contrast). Among the SIPs found by Lucas-Kanade, there are many of them located at "non-intent" regions, say not only at the edge/boundary, but also at the regions that have small brightness contrast. However, we should not be too disappointed by the visual performance of Lucas-Kanade detector. Even though it is not competitive enough against eight-way or Harris when it is individually used in case studies, it has been found to be helpful in improving the classification accuracies when the SIPs found by Lucas-Kanade are combined with eight-way or Harris, or both (by concatenating their distributions).

We can observe that the SIPs (of KA and KL) found by eight-way or Harris are mostly distributed around the facial features (say eye, nose, etc). Comparing the SIPs distributions in columns (between different classes), we will have the expectation that the first two pairs are more discriminant than the third pair (contributed by Lucas-Kanade).

4 Distribution of SIPs

A SIP distribution (of an image) represents the number of occurrence of SIPs in each fixed small block of an image. Using regular (rectangle) grids is a simple method to split an image region in multiple blocks.

There are two important parameters in order to build a good SIP distribution. One parameter is the number of SIPs to be used, and the other is the size of rectangle grids. For a general view on distribution construction, more SIPs leads to more accurate distribution. But in practice, we may need some compromise between the accuracy and computation cost. In our study, we found that 300 SIPs of an image are usually sufficient to build a "good" distribution representation (500 SIPs can contribute slightly better representation).

Another important parameter is the size of rectangle grids. A basic principle in deciding the size of rectangle grids is that: *it should be small enough to accurately characterize the local information*. We use 4×8 as the size of rectangle grids (except for Yale dataset where 16×16 is used). Remind that an image plane is usually a rectangle. The distribution representation is not invariant to image translation/video temporal shift. However, if the visual media data is aligned, SIP distribution can then faithfully represent its essential content/information. The face images in many datasets are aligned, thus can be used directly as our case studies.

We now complete the discussion of SIP distribution with its algorithm implementation:

Algorithm 3 Distribution of SIPs: get_dist()

```
Input: SIPs of one image (obtained from some SIP miner)
Output: SIP distribution: dist
  1. Initialize the grid size $(g_1, g_2)$
  2. For each SIP $p = (p_x, p_y)$
     2.1 dist($[p_x/g_1]$, $[p_y/g_2]$)++
  3. Align the dist from a matrix to a vector
```

Fig. 4 shows the feature vectors of the face images of KA and KL (given in Fig. 3). Consistently with the visual measurement, the SIP distribution contributed by eight-way has richer "structure", whereas, the structure of Lucas-Kanade is the flattest one.

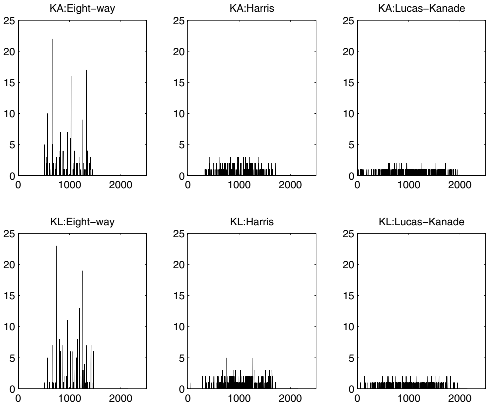

Figure 4. The SIP distributions of eight-way, Harris, and Lucas-Kanade of the face images, KA and KL.

5 Class-preserving dimension reduction

A general dimension reduction problem is formulated as follows: Given a data matrix $A \in R^{m\times n}$, where each column corresponds to a data point, to find a linear transformation $G^T \in R^{\ell\times m}$ that maps each column a_i, for $1 \leq i \leq n$, of A in the m dimensional space to a column y_i in the ℓ dimension space:

$$G^T : a_i \in R^{m\times 1} \to y_i \in R^{\ell\times 1}. \tag{5.3}$$

In this section, we will study the class-preserving dimension reduction. The class information is known and the data matrix A is formulated by: $A = [A_1 \quad A_2 \quad \cdots \quad A_k]$, where k is the number of classes, and $\sum_{i=1}^{k} n_i = n$, and $A_i \in R^{m\times n_i}$, is collection of data points in i-th class. We will first have a brief review on classic LDA, and then formulate an objective/energy function for reduction transformation G^T using the ratio of the traces of within-class scatter, and between-class scatter in low dimension space. Similar work has been done in [22], where no explicit global objective function is defined.

5.1 Classic LDA

In classic LDA, two scatter matrices, within-class scatter matrix S_w, and between-class scatter matrix S_b are defined to quantify the quality of the classes, as follows,

$$S_w = H_w H_w^T, \quad S_b = H_b H_b^T. \tag{5.4}$$

where

$$\begin{aligned} H_w &= [A_1 - c^{(1)}(e^{(1)})^T, \cdots, A_k - c^{(k)}(e^{(k)})^T] \in R^{m\times n}, \\ H_b &= [\sqrt{n_1}(c^{(1)} - c), \cdots, \sqrt{n_k}(c^{(k)} - c)] \in R^{m\times k}, \end{aligned} \tag{5.5}$$

and the centroid $c^{(i)}$ of the ith class is defined as $c^{(i)} = \frac{1}{n_i} A_i e^{(i)}$ where $e^{(i)} = (1, 1, \cdots, 1)^T \in R^{n_i \times 1}$, and the global centroid c is defined as $c = \frac{1}{n} Ae$ where $e = (1, 1, \cdots, 1)^T \in R^{n\times 1}$.

Classic LDA finds transformation G^T as eigenvectors (associated with the smallest eigenvalues) of matrix $S_b^{-1} S_w$, which requires S_b to be nonsingular.

5.2 Dimension reduction by optimizing trace ratio

A way to overcome the requirement of nonsingular scatters in classic LDA is using the trace ratio. Assume N_i be the set of column indices that belong to the ith class. Let us first have a look at what are the traces of within-class and between-class scatter matrices of original data:

$$\begin{aligned} \text{trace}\,(S_w) &= \sum_{i=1}^{k} \sum_{j\in N_i} (a_j - c^{(i)})^T (a_j - c^{(i)}) \\ &= \sum_{i=1}^{k} \sum_{j\in N_i} ||a_j - c^{(i)}||^2 \qquad (5.6) \\ \text{trace}\,(S_b) &= \sum_{i=1}^{k} n_i (c^{(i)} - c)^T (c^{(i)} - c) \\ &= \sum_{i=1}^{k} n_i ||c^{(i)} - c||^2. \qquad (5.7) \end{aligned}$$

So trace(S_w) and trace(S_b) characterize the closeness of the points within a class, and the separation between classes separately. Thus, small trace(S_w) and large trace(S_b) are desirable in order to achieve good classification rates in real applications.

Denote $S_w^L = (G^T H_w)(G^T H_w)^T = G^T H_w H_w^T G = G^T S_w G$. Similarly, we have the between-class covariance matrices $S_b^L = G^T S_b G$. The goal of a class-preserving dimension reduction is thus to find a reduction transformation G^T to minimize the ratio of the traces of scatter matrices S_w^L and S_b^L, i.e.,

$$\min F(G) = \frac{\text{trace}(S_w^L)}{\text{trace}(S_b^L)}. \tag{5.8}$$

We can solve the trace optimization problem (5.8) using the generalized singular value decomposition (GSVD) [14]. Denote $K = [H_b, H_w]$. The GSVD on the matrix pair (H_b^T, H_w^T), will give orthogonal matrices $U \in R^{k\times k}$, $V \in R^{n\times n}$, and a nonsingular matrix $X \in R^{m\times m}$, such that

$$\begin{bmatrix} U & 0 \\ 0 & V \end{bmatrix}^T KX = \begin{bmatrix} \Sigma_1 & 0 \\ \Sigma_2 & 0 \end{bmatrix}. \tag{5.9}$$

where $\Sigma_1 = \begin{bmatrix} I_b & 0 & 0 \\ 0 & D_b & 0 \\ 0 & 0 & 0_b \end{bmatrix}$, and $\Sigma_2 = \begin{bmatrix} 0_w & 0 & 0 \\ 0 & D_w & 0 \\ 0 & 0 & I_w \end{bmatrix}$. Here $I_b \in R^{r\times r}$ is an identity matrix with $r = \text{rank}(K) - \text{rank}(H_w^T)$, $D_b = \text{diag}(\alpha_{r+1}, \cdots, \alpha_{r+s})$, and $D_w = \text{diag}(\beta_{r+1}, \cdots, \beta_{r+s}) \in R^{s\times s}$, are diagonal matrices with $s = \text{rank}(H_b) + \text{rank}(H_w) - \text{rank}(K)$, satisfying $1 > \alpha_{r+1} \geq, \cdots, \geq \alpha_{r+s} > 0$, $0 < \beta_{r+1} \leq, \cdots, \leq \beta_{r+s} < 1$, and $\alpha_i^2 + \beta_i^2 = 1$ for $i = r+1, \cdots, r+s$.

Denote u_{ii} is the ii-th term of the matrix $X^{-1}GG^TX^T$. Trace optimization problem (5.8) can be re-formulated as the following problem:

$$\begin{aligned} &\text{minimize} \quad \text{trace}(S_w^L) = \sum_{i=1}^{t} u_{ii} - 1 \\ &\text{subject to} \quad \text{trace}(S_b^L) = \sum_{i=1}^{r} u_{ii} + \sum_{i=r+1}^{r+s} \alpha_i^2 u_{ii} = 1. \end{aligned} \tag{5.10}$$

We now present a novel result in Thm 5.1. Limited by the space, we refer the readers to our technical report [28] for the details of proof. A simple algorithm to compute GSVD can be found in [22], based on [21]. GSVD is also provided in Matlab software. Our class-preserving dimension reduction algorithm is summarized in Algorithm 4. The time/space complexity of PCA and LDA/GSVD is essentially the same to each other, as shown in Table 1. More details can be found in [28].

Theorem 5.1 *If $\{u_{ii}^\star\}_{i=1}^m$ is an optimal solution of the optimization problem (5.10), then $u_{ii}^\star \geq u_{jj}^\star$, for $1 \leq i \leq j \leq r+s$.*

Algorithm 4 GSVD based dimension reduction

```
Input: High dimension feature vectors of training data
Ouput: Reduction transformation $G^T$
Main variables:
    $H_b$ - between-class covariance matrix
    $H_w$ - within-class covariance matrix
    $G^T$ - reduction transformation
1. Construct the matrices $H_b$ and $H_w$ as defined in (5.5)
2. Compute GSVD on the matrix pair $(H_b^T, H_w^T)$, and get the matrix $X$ as in (5.9).
3. Compute rank($H_b$) and assign it to $u$
4. Let $X_u^T$ be the first $u$ columns of $X$
5. Assign transformation matrix $G^T = X_u^T$
```

Methods	Time Complexity	Space Complexity
PCA	$O(m^2 n)$	$O(nm)$
LDA/GSVD	$O((m+k)^2 n)$	$O(nm)$

Table 1. Complexity comparison: n is the number of training data points, m is the number of the dimensions, and k is the number of classes.

6 Classifier: nearest neighbor

Nearest neighbor classifier is used in our case studies, which is quite simple. Given a set of training data points (that are labeled) and a query data point, we compute the distance (or similarity) between the query data point to each training points. The query data point is annotated as the same class label as the one which has the shortest distance to the query point.

Given two data points, $q = (q_1, q_2, \cdots, q_d)$ and a point $q = (q'_1, q'_2, \cdots, q'_d)$ in R^d, their distance can be measured by the L_p-norm, i.e., $L_p(q, q') = \left[\sum_{i=1}^{d} |q_i - q'_i|^m\right]^{1/p}$. L_2-norm (i.e., Euclidean distance) is the most widely used metric because of its convenient analytic properties (note that the GSVD-based class-preserving dimension reduction assumes Euclidean distance because of the trace optimization). However the robust statistic literature shows that L_2 overly penalizes outliers [10]. So in our case studies, we will try different L_p-norms on the space of SIP distributions, whereas apply L_2-norm in the Eigenspace or Fisherspace[8]. Our experiments will show that $p < 1$ does outperform Euclidean distance. Using the sub-sample pixel values of images as feature vector, [24] also observed the superiority of $p < 1$ in face recognition task.

7 Case studies

In this section, we present two case studies on the visual media data based classification using SIP feature vectors. They are static facial expression recognition, and face recognition. 10-fold cross validation method is used to estimate the classification accuracy.

The results on one dataset are presented in one table. To present the results compactly, we will use the following simplified notations: e = eight-way SIP distribution, shortly e = eight-way; Similarly, h = Harris; l = Lucas-Kanade; e+h = eight-way concatenated by Harris; e+l = eight-way concatenated by Lucas-Kanade; h+l = Harris concatenated by Lucas-Kanade; all = the concatenation of three distributions. The values appearing in the first row of tables are the norm index for nearest neighbor classifier. The first column of the result tables will be indexed by these simplified notations. The last two columns are the classification accuracy (in percentage) based on the high-level features extracted by the PCA dimension reduction and our class-preserving dimension reduction. They are used only in Euclidean space because trace optimization is based on L_2-norm. With the left three columns, we can analyze what norm is good at measuring SIP distributions. With the right two columns, we can analyze how well the dimension reduction techniques work.

Each table will present the answers for the following four questions:

1. What is the best SIP miner (by comparing the best accuracies that they can reach in individual use)?
2. Which L_p-norm performs the best?
3. What is the best accuracy?
4. How well do PCA and the class-preserving dimension reduction (denoted by C-P) work?

For the convenience of reading, we highlight/underline the answers for the first three questions in each table. The overall conclusion on all experiments will be given the next section.

7.1 Datasets

We have one dataset for universal facial expression recognition, and four datasets for face recognition. Jaffe was originally built for the study of static facial expression recognition [18], but will also be used for face recognition in our case study. There are little variation in translation, pose, occlusion and lighting. Jaffe contains 10 female faces, 7 universal facial expressions, and all together 210 image instances. So when it is used as a dataset for facial expression recognition, each class has 30 instances; when it is used as a dataset for face recognition, each class has 21 instances. In either application context, there are always some image instances visually very similar to each other, which objectively creates the opportunity of achieving high accuracy on this dataset.

The images in Yale face dataset have the variations in facial expressions (as well as Jaffe) in addition to illumination. The publicly available Yale face dataset is aligned to some degree but not perfectly. Yale dataset contains 15

[8] Eigenspace and Fisherspace refers to the reduced spaces via PCA and LDA (either classic or generalized) respectively.

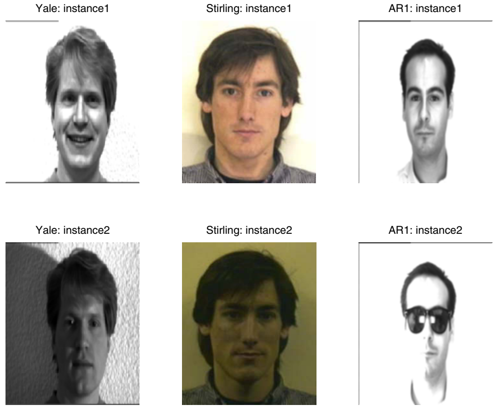

Figure 5. Yale, Stirling, AR1 each have two samples

classes/faces and each face has 11 instances. The major variation in Stirling face images is pose and color. Stirling contains 659 face images. We use the first 100 images (of 10 classes) in our case studies. The last face dataset we will test is named AR whose images contains the variations in occlusion and facial expressions. AR is a huge dataset of face images. We use the first 100 images (of 9 classes)in our case studies.

Fig. 5 shows some examples of the face image datasets, Yale, Stirling and AR1. The dimensions of the SIP feature vectors of these four datasets are 2048, 336, 1100 and 3577 respectively. Note that LDA (either classic or generalized) always reduces the original dimensionality to the number of classes minus 1. Thus The reduced dimensions in our case studies are 9, 14, 9, and 8 respectively.

7.2 Case study 1: Static universal facial expression recognition

Tab.2 shows the classification accuracies on Jaffe as a facial expression dataset. Without using the class information of training data, PCA was found to severely degrade the accuracies by using the SIP distributions, whereas C-P can approximate the accuracies very well.

7.3 Case study 2: Face recognition

Tab. 3, 4, 5 and 6 show the classification accuracies on Jaffe, Yale, Stirling and AR1 dataset respectively. PCA usually degrades the accuracy downfrom 15% to 25%, whereas the class-preserving dimension reduction preserves (either slightly degrades or slightly increases) the accuracies by using SIP distributions.

8 Conclusion and future work

In this paper, we present a comparative study on the spatial interest pixels (SIPs). With extensive experiments, we

SIPs	1/2	3/4	1	2	2	2
					PCA	C-P
e	81.2	81.6	82.0	81.2	60.2	79.5
h	78.0	79.5	78.3	77.4	54.6	77.6
l	67.2	68.0	68.2	67.0	47.3	65.2
e+h	89.1	90.6	91.5	89.1	66.2	86.2
e+l	81.5	82.3	83.0	82.3	61.4	80.5
h+l	88.5	90.2	91.5	89.1	66.2	86.2
all	89.2	91.4	**91.9**	90.0	66.7	87.6

Table 2. Classification accuracy on Jaffe as a facial expression recognition dataset.

SIPs	1/2	3/4	1	2	2	2
					PCA	C-P
e	99.5	99.5	98.1	94.8	78.3	98.6
h	96.2	99.5	99.5	99.5	84.2	99.1
l	99.1	99.1	98.6	98.6	81.6	98.1
e+h	99.5	99.5	99.5	96.7	79.2	99.1
e+l	99.5	99.5	99.1	96.2	80.5	99.5
h+l	99.5	99.5	99.5	99.5	82.3	98.6
all	**99.5**	**99.5**	**99.5**	**99.5**	83.8	99.1

Table 3. Classification accuracy on Jaffe as face recognition dataset.

SIPs	1/2	3/4	1	2	2	2
					PCA	C-P
e	88.7	90.0	90.7	85.3	82.5	82.0
h	90.7	91.3	91.3	88.6	86.0	90.0
l	85.3	85.3	87.3	89.0	85.3	83.3
e+h	94.7	94.8	94.7	89.0	86.0	91.3
e+l	94.0	95.3	95.3	89.3	88.7	92.7
h+l	94.7	94.7	94.1	92.7	90.7	91.3
all	95.3	**96.0**	94.7	90.0	88.0	93.3

Table 4. Classification accuracy on YALE dataset as face recognition.

SIPs	1/2	3/4	1	2	2	2
					PCA	C-P
e	93.8	95.0	93.8	86.3	78.2	86.3
h	88.8	87.5	85.0	83.8	75.2	92.5
l	60.0	62.5	61.2	56.2	48.5	61.2
e+h	93.8	95.0	95.0	93.8	85.0	95.0
e+l	93.8	95.0	93.8	86.3	79.2	93.8
h+l	81.3	81.3	82.5	82.5	74.8	86.2
all	93.8	**95.0**	**95.0**	93.8	85.0	95.0

Table 5. Classification accuracy on Stirling sub dataset (first 100 images 10 classes).

SIPs	1/2	3/4	1	2	2	2
					PCA	C-P
e	92.0	95.0	92.0	87.0	69.0	93.0
h	90.0	93.0	93.0	93.0	72.0	92.0
l	86.0	88.0	87.0	83.0	65.0	85.0
e+h	94.0	94.0	95.0	95.0	74.0	95.0
e+l	94.0	95.0	96.0	89.0	68.0	93.0
h+l	96.0	96.0	96.0	93.0	72.0	95.0
all	**98.0**	**98.0**	96.0	96.0	70.0	95.0

Table 6. Classification accuracy on AR1 (first 100 instances, 8 classes).

have shown that SIPs are useful low-level features for visual media data. The best classification accuracies on these applications we can achieve are either higher than or close to those in the literature.

With the comparative study on different applications, we have the following overall conclusions on the use of SIP features (mainly on SIP distributions):

- Eight-way SIP miner is statistically the best. Among the 5 experiments/Tables, eight-way wins 3 times, ties one time (with Harris), and loses one time
- SIP distributions contributed by different SIP miner can be concatenated together into longer feature vectors that are more useful for different applications.
- The best distance measure in SIP feature vector space is L_p-norm with $p < 1$, In static facial expression recognition and face recognition, $p = 0.75$ are usually the "best" (among the four options provided in this paper).
- The GSVD-based dimension reduction can essentially preserve the classification accuracy on SIP feature vectors. It distinctly outperforms PCA dimension reduction.

In the future, we will study spatial-temporal interest pixel to reduce the dimensionality of video data.

References

[1] S. Arya. Nearest neighbor searching and applications. In *Ph. D. Thesis, University of Maryland, College Park, MD*, 1995.

[2] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE TPAMI*, 19(7):711–720, 1997.

[3] J. Bergen and M. Landy. Computational modeling of visual texture segregation. In *Computational Models of Visual Perception*, pages 253–271. MIT Press, Cambridge MA, 1991., 1991.

[4] R. Chellappa, C. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5):705–740, 1995.

[5] P. Ekman and W. Friesen. Pictures of facial affect. In *Consulting psychologist, Palo Alto, CA*, 1976.

[6] R. Fisher. The use of multiple measurements in taxonomic problems. In *Annals of Eugenics 7*, pages 179–188, 1936.

[7] T. Gevers and A. W. M. Smeulders. Image indexing using composite color and shape invariant features. In *ICCV*, pages 576–581.

[8] P. Hancock, A. Burton, and V. Bruce. Face processing : Human perception and principal components analysis. In *Memory Cognition, 24*, pages 26–40, 1996.

[9] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conference, Manchester*, pages 147–151, 1988.

[10] P. Huber. *Robust Statistics*. Wiley, 1981.

[11] I. Jolliffe. Principle component analysis. *Journal of Educational Psychology*, 24:417–441, 1986.

[12] D. Joyce, P. lewis, R. Tansley, M. Dobie, and W. Hall. Semiotics and agents for integrating and navigating through multimedia representations of concepts. In *Proceedings of SPIE Vol. 3972, Storage and Retrieval for Media Databases 2000*, pages 132–143, 2000.

[13] W.-H. Lin and A. Hauptmann. News video classification using svm-based multimodal classifiers and combination strategies. In *ACM Multimedia, Juan-les-Pins, France*, 2002.

[14] C. V. Loan. Generalizing the singular value decomposition. *SIAM Journal on Numerical Analysis*, 13(1):76–83, 1976.

[15] E. Loupias and N. Sebe. Wavelet-based salient points for image retrieval. In *RR 99.11, Laboratoire Reconnaissance de Formes et Vision, INSA Lyon*, November 1999.

[16] Y. Lu, C. Hu, X. Zhu, H. Zhang, and Q. Yang. A unified framework for semantics and feature based relevance feedback in image retrieval systems. In *ACM Multimedia*, pages 31–37, 2000.

[17] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[18] M. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE transcations on PAMI*, 21(12):1357–1362, 1999.

[19] A. Martinez and R. Benavente. The ar face database. Technical Report CVC Tech. Report No. 24, 1998.

[20] A. Martinez and A. Kak. PCA versus LDA. *IEEE TPAMI*, 23(2):228–233, 2001.

[21] C. Paige and M.A.Saunders. Towards a generalized singular value decomposition. *SIAM Journal on Numerical Analysis*, 18:398–405, 1981.

[22] H. Park, P. Howland, and M. Jeon. Cluster structure preserving dimension reduction based on the generalized singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, to appear.

[23] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000.

[24] T. Sim, R. Sukthankar, M. Mullin, and S. Baluja. Memory-based face recognition for visitor identification. In *Proc. 4th Intl. Conf. on FG'00*, pages 214–220.

[25] J. Smith. Integrated spatial and feature image systems: Retrieval and compression. In *PhD thesis, Graduate School of Arts and Sciences, Columbia University, New York, NY*, 1997.

[26] M. Swain and D. Ballard. Color indexing. *Int. J. computer vision*, 7:11–32, 1991.

[27] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[28] J. Ye, R. Janardan, C. park, and H. Park. A new optimization criterion for generalized discriminant analysis on undersampled problems. Technical Report TR-026-03, 2003.

[29] Z. Zhang. Feature-based facial expression recognition: experiments with a multi-layer perceptron. *Int.l Journal of Pattern Recognition and Artificial Intelligence*, 13(6):893–911, 1999.

[30] W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips. Face recognition: A literature survey. Technical Report CAR-TR-948, 2000.