Principal motion: PCA-based reconstruction of motion histograms

Hugo Jair Escalante[a] and Isabelle Guyon[b]
[a] INAOE, Puebla, 72840, Mexico,
[b] CLOPINET, Berkeley, CA 94708, USA
`http://chalearn.org`

May 2012

`principal_motion` is the implementation of a reconstruction approach to gesture recognition based on principal components analysis (PCA). The underlying idea is to perform PCA on the frames in each video from the vocabulary, storing the PCA models. Frames in test-videos are projected into the PCA space and reconstructed back using each of the PCA models, one for each gesture in the vocabulary. Next we measure the reconstruction error for each of the models and assign a test video the gesture that obtains the lowest reconstruction error. The rest of this document provides more details about the `principal_motion` object.

1 Representation

Let $\mathcal{V}$ be a video composed of N frames, $\mathcal{V} = \{I_1, \ldots, I_N\}$, where I_i is the i^{th} frame, each frame is a grayscale image. Let $\mathbb{V} = \{\mathcal{V}_1, \ldots, \mathcal{V}_K\}$ be the set of videos corresponding the the gesture vocabulary, where each video corresponds to a gesture.

A video is represented by a set of histograms of motion energy $H_{1,\ldots,N-1}$ obtained from difference images as follows. First, we obtain motion energy information by subtracting consecutive frames (from frames 1 to $N-1$) in the video: $D_i = I_{i+1} - I_i$, $i = \{1, \ldots, N-1\}$, thus the video $\mathcal{V}$ is associated to $N-1$ difference images, which are then processed to obtain motion histograms. Figure 1 shows a difference image obtained from a sample video.

A grid of equally spaced cells (bins) is defined over the difference image being processed. The size of the grid is denoted by ρ and it is the same

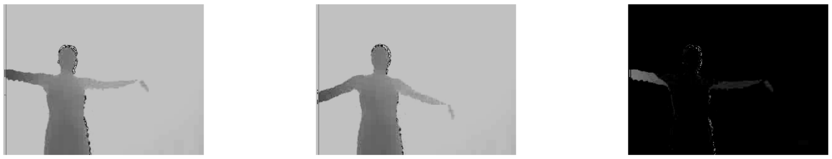

Figure 1: A difference image (right), obtained by subtracting two consecutive frames (left, center).

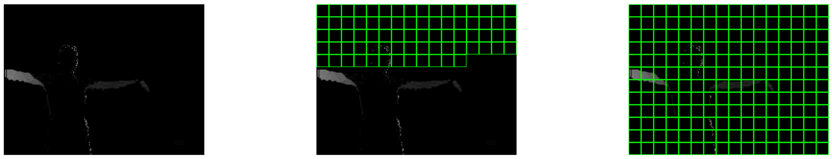

Figure 2: Grid associated to a difference image.

for all of the images, see Figure 2. We estimate for each difference image $D_i \in \{D_1, \ldots, D_{N-1}\}$, the average motion energy in each of the cells of the grid. That is, we obtain a 2D histogram for each difference image, where each bin accounts for the average motion energy in the image and the corresponding 2D bin. The 2D histograms are transformed into a vector of dimensions $1 \times N_b$, where Nb is the number of bins in the grid. Thus a video $\mathcal{V}_i$ is represented by a matrix $\mathbf{Y}_i$ of dimensions $(N-1) \times N_b$, with one row per difference image and one column per 2D bin. Each video is processed and represented in this way.

2 Recognition

For recognition we consider a reconstruction approach based on PCA. We apply PCA (via singular value decomposition) to each of the matrices $\mathbf{Y}_1, \ldots, \mathbf{Y}_K$ that represent videos in $\mathbb{V}$ (i.e., the gesture vocabulary). We store the top c eigenvalues, $\mathbf{W}$, and the corresponding eigenvectors, $\mathbf{U}$. Hence we have for each gesture in the vocabulary a PCA model represented by the pair $(\mathbf{W}, \mathbf{U})_{\{1,\ldots,K\}}$. Figure 3 shows the top$-9$ principal components of a specific gesture.

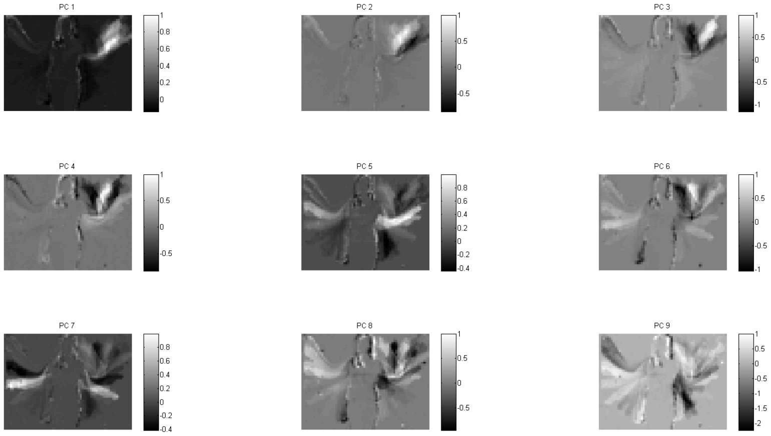

Figure 3: First 9 principal components of a particular gesture.

When an unseen video $\mathcal{T}$ needs to be classified it is segmented[1] so that a single gesture is contained in the video. $\mathcal{T}$ is processed similarly as training videos and hence it is represented by a matrix of motion histograms $\mathbf{T}$. Matrix $\mathbf{T}$ is projected into each of the $K-$ spaces induced by $(\mathbf{W}, \mathbf{U})_{\{1,\ldots,K\}}$, and projections are reconstructed back. Let $\mathbf{R}_1, \ldots \mathbf{R}_K$ denote the reconstructions of $\mathbf{T}$ according to PCA models $1, \ldots, K$, respectively. We measure the reconstruction error for each $\mathbf{R}_i$ as follows: $\epsilon(h) = \frac{1}{n} \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} (\mathbf{R}_{i,j} - \mathbf{T}_{i,j})^2}$, where n and m are the number of rows and columns of $\mathbf{T}$, respectively and with $h = 1, \ldots, K$. That is, we measure the average reconstruction error over the motion histograms associated to the video $\mathcal{T}$. Finally, we assign $\mathcal{T}$ the gesture corresponding to the PCA model that obtained the lowest reconstruction error, that is: $\arg\min_h (\epsilon(h))$, see Figura 4.

The PCA reconstruction approach to gesture recognition is inspired from the one-class classification task, where the reconstruction error via PCA has been used to identify outliers [2]. The method is also inspired in a recent method for spam classification [1]. The underlying hypothesis of the method is that a test video will be better reconstructed with a PCA model that was obtained with another video that contains the same gesture.

[1] We use a simple method for temporal segmentation based on the average gesture duration in the training set.

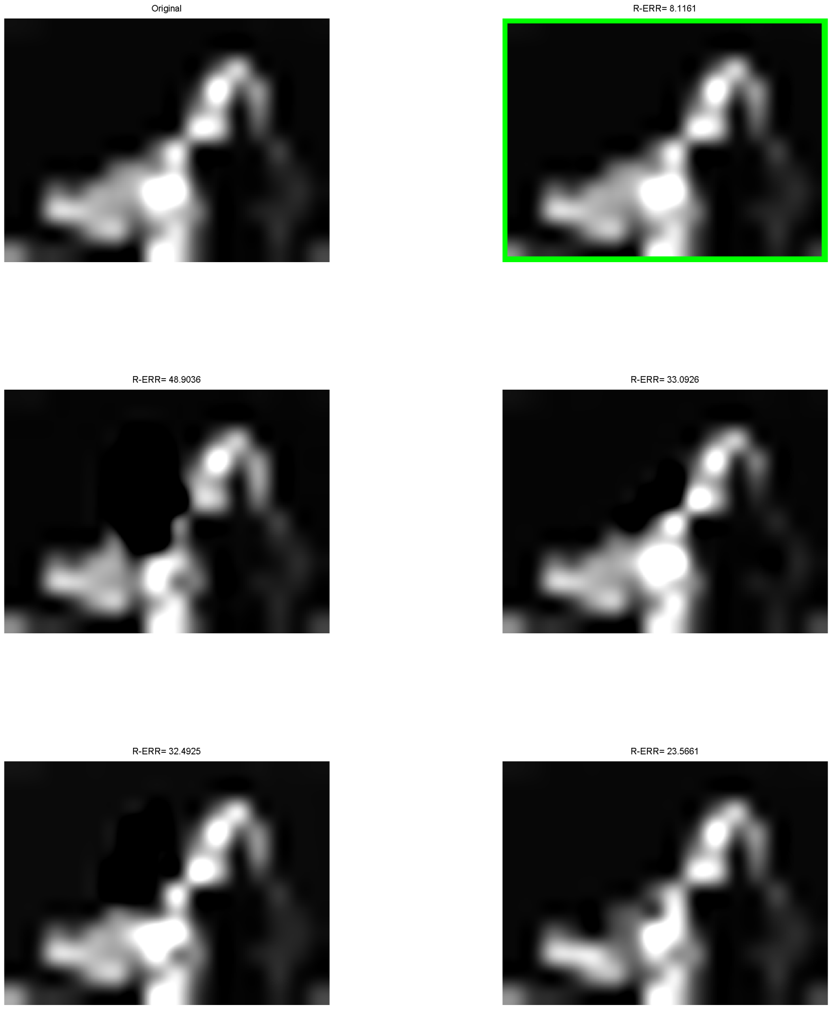

Figure 4: Cumulative images obtained by adding all of the reconstructed frames of a video using different PCA models. The top-left image shows the original video, and the highlighted image shows the reconstruction with the PCA model associated to the correct gesture. The other images show reconstructions of the original video generated with different (incorrect) PCA models. Above each image we show the reconstruction error obtained with the corresponding PCA model.

References

[1] J. C. Gomez and M. F. Moens. Pca document reconstruction for email classification. *Computational Statistics and Data Analysis*, 56:741–751, 2012.

[2] D. Tax. *One-Class Classification*. PhD thesis, Delft University of Technology, 2001.