Infinite Gaussian Fisher Vector to Support Video-Based Human Action Recognition

Jorge L. Fernández-Ramírez[1(✉)], Andrés M. Álvarez-Meza[2], Álvaro A. Orozco-Gutiérrez[1], and Julian David Echeverry-Correa[1]

[1] Automatics Research Group, Universidad Tecnológica de Pereira, Pereira, Colombia
jorgeferram17@utp.edu.co
[2] Signal Processing and Recognition Group, Universidad Nacional de Colombia - Sede Manizales, Manizales, Colombia

Abstract. Human Action Recognition (HAR) is a computer vision task that attempts to monitor, understand, and characterize humans in videos. Here, we introduce an extension to the conventional Fisher Vector encoding technique to support this task. The methodology, based on the Infinite Gaussian Mixture Model (IGMM) seeks to reveal a set of discriminant local spatio-temporal features for enabling the precise codification of visual information. Specifically, it is much simpler to handle the infinite limit from the IGMM, than working with traditional Gaussian Mixture Models (GMMs) with unknown sizes, that will require extensive cross-validation. Under this premise, we developed a fully automatic encoding methodology that avoids heuristically specifying the number of components in the mixture model. This parameter is known to greatly affect the recognition performance, and its inference with conventional methods implies a high computational burden. Moreover, the Markov Chain Monte Carlo implementation of the hierarchical IGMM effectively avoids local minima, which tend to plague mixtures trained by optimization-based methods. Attained results on the UCF50 and HMDB51 databases demonstrate that our proposal outperforms state of the art encoding approaches concerning the trade-off between recognition performance and computational complexity, as it drastically reduces both number of operations and memory requirements.

Keywords: Human Action Recognition · Infinite Gaussian Mixture Model · Fisher Vector · Video processing

1 Introduction

Human action recognition (HAR) is a computer vision task that seeks to monitor, understand, and characterize humans in videos [7]. This task has a wide pool of applications that include automatic surveillance, video indexing and retrieval, and virtual reality [5]. The conventional pipeline for action recognition can be

G. Bebis et al. (Eds.): ISVC 2019, LNCS 11845, pp. 38–49, 2019.
https://doi.org/10.1007/978-3-030-33723-0_4

divided into three stages: (i) feature extraction from raw videos, (ii) data representation, (iii) and classification into predefined categories [20].

In regard to feature extraction, the literature exhibits two trends, hand-crafted and Convolutional Neural Networks (CNNs) features. Both intend to describe the local space, codify the motion information, and then combine these sources for allowing the proper transcription of human activity [15]. To the date, Two-stream CNNs is the most effective framework for action recognition, employing two deep networks and fusion techniques to take advantage of both appearance and motion clues [13]. However, CNN-based methods hamper in-depth action analysis and understanding as there is not visual interpretability [22]. Moreover, deep learning requires large amounts of training data, which in many applications is not available [1]. On the contrary, The most popular hand-crafted feature estimation technique is known as Improved Dense Trajectories (iDT) [20]. The method, describes the local space of trajectories generated by tracking a dense grid of points. Employing descriptors such as Histograms of Oriented Gradients (HOG) for codifying appearance through color gradients, Histograms of Optical Flow (HOF) for describing movement, and Motion Boundary Histograms (MBH) for codifying changes in motion [3].

For data representation, authors proposed feature encoding and relevance analysis for highlighting salient patterns and enabling codification of visual information [12]. Super-vector based methods such as Fisher Vector (FV) and Vector of Locally Aggregated Descriptors (VLAD) are presented as the most well-known approaches for feature encoding in action recognition tasks [19]. On the other hand, non-linear relevance analysis using kernel methods have shown promising results in recent research [7]. Nevertheless, their kernel evaluation requires computing and storing large distance matrices, while also tuning parameters which increases computational complexity [7]. Lastly, it is convention to employ Support Vector Machines (SVM) for classification [17].

Both FV and VLAD methods are supported by the Gaussian Mixture Model (GMM) to generate a codebook of visual words [21]. These methods quantify the similarity between a video sample and previously computed codebook for encoding visual information through calculating Gaussian responsibilities [6]. However, GMMs trained by optimization-based methods, e.g. Expectation Maximization (EM), require extensive cross-validation for selecting the number of visual words in the codebook [8]. Moreover, the initialization required by these training methods makes models fall into local minima [2]. Therefore, using conventional GMM implies large number of operations and memory requirements, which increases the computational burden of conventional recognition systems.

In this paper, we introduce a novel data encoding framework using Bayesian inference and Dirichlet processes to support video-based HAR. Our approach is fully automatic, allowing every parameter in the model to be updated hierarchically through the Markov Chain Monte Carlo (MCMC) algorithm Gibbs sampling. Specifically, our approach includes a Infinite Gaussian Mixture model (IGMM) for revealing a set of discriminant visual words, trained through a MCMC-based optimization that evades local minima. In Fact, the infinite limit

on the number components avoids estimating this parameter through extensive cross-validation. Attained results on both UCF50 and HMDB51 databases demonstrate that our proposal obtained promising recognition performance and computational savings, favoring HAR tasks.

The rest of the paper is organized as follows: Sect. 2 presents the main theoretical background. Section 3 describes the experimental setup. Section 4 introduces results and discussions. Finally, Sect. 5 presents conclusions and future work.

2 Infinite Gaussian Model for Fisher Vector Encoding

Let $\{\boldsymbol{Z}_n \in \mathbb{R}^{T_n \times D}, y_n \in \mathbb{N}\}_{n=1}^{N}$ be an input-output pair set holding N human action videos. Each sample $\boldsymbol{Z}_n$, is represented by T_n observations. The local space of every observation is characterized by a D-dimensional descriptor, as in [20]. The output label y_n denotes the specific human action of video n. From $\boldsymbol{Z} \in \mathbb{R}^{T \times D}$, where $T = \sum_{n=1}^{N} T_n$, we aim to train a generative model using IGMM. The procedure is as follows [4]:

The likelihood from observation $\boldsymbol{z}_t \in \mathbb{R}^D$ to a GMM with k_{rep} components is:

$$p(\boldsymbol{z}_t | \{\boldsymbol{\mu}_j, \boldsymbol{S}_j, \pi_j\}_{j=1}^{k}) = \sum_{j=1}^{k_{\text{rep}}} \pi_j \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{S}_j^{-1}) \tag{1}$$

where $\boldsymbol{\mu}_j \in \mathbb{R}^D$ are mean vectors, $\boldsymbol{S}_j \in \mathbb{R}^{D \times D}$ are precision matrices, and π_j are the mixing proportions. Variable k_{rep} denotes the number of Gaussian components that have associated data, named represented classes [14].

2.1 Component Parameters

The component means $\boldsymbol{\mu}_j$ and precisions $\boldsymbol{S}_j$ are given by Gaussian and Wishart priors, respectively:

$$p(\boldsymbol{\mu}_j | \boldsymbol{\lambda}, \boldsymbol{R}) \sim \mathcal{N}(\boldsymbol{\lambda}, \boldsymbol{R}^{-1}) \qquad p(\boldsymbol{S}_j | \beta, \boldsymbol{W}) \sim \mathcal{W}(\beta, \boldsymbol{W}^{-1}) \tag{2}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^D$ is a mean vector, $\boldsymbol{R} \in \mathbb{R}^{D \times D}$ and $\boldsymbol{W} \in \mathbb{R}^{D \times D}$ are precision matrices, and β is the degrees of freedom. These hyper-parameters are common to all components. The conditional posterior on $\boldsymbol{\mu}_j$ is obtained by conjugating its prior:

$$\begin{aligned} p(\boldsymbol{\mu}_j | \boldsymbol{\lambda}, \boldsymbol{R}, \{\boldsymbol{z}_t : c_{t,j} = 1\}, \boldsymbol{S}_j) &\propto \prod_{t:c_{t,j}=1} p(\boldsymbol{z}_t | \boldsymbol{\mu}_j, \boldsymbol{S}_j) \times p(\boldsymbol{\mu}_j | \boldsymbol{\lambda}, \boldsymbol{R}) \\ &\sim \mathcal{N}\left((T_j \, \overline{\boldsymbol{z}}_j \, \boldsymbol{S}_j + \boldsymbol{\lambda}\boldsymbol{R})(T_j \, \boldsymbol{S}_j + \boldsymbol{R})^{-1}, (T_j \, \boldsymbol{S}_j + \boldsymbol{R})^{-1}\right) \end{aligned} \tag{3}$$

$\boldsymbol{c}_t \in \mathbb{R}^k$ is a latent variable, with notation 1 of k, where k include both represented and unrepresented classes. Unrepresented classes are virtually infinite [18]. T_j is

the number of observations belonging to class j. likewise, $\overline{\boldsymbol{z}}_j$ is the average vector of these observations.

$$\overline{\boldsymbol{z}}_j = \frac{1}{T_j} \sum_{t:c_{t,j}=1} \boldsymbol{z}_t, \quad T_j = \sum_{t=1}^{T} c_{t,j} \tag{4}$$

The conditional posterior on $\boldsymbol{S}_j$ is obtained by conjugating its prior:

$$\begin{aligned} p(\boldsymbol{S}_j|\beta, \boldsymbol{W}, \{\boldsymbol{z}_t : c_{t,j}=1\}, \boldsymbol{\mu}_j) &\propto \prod_{t:c_{t,j}=1} p(\boldsymbol{z}_t|\boldsymbol{\mu}_j, \boldsymbol{S}_j) \times p(\boldsymbol{S}_j|\beta, \boldsymbol{W}) \\ &\sim \mathcal{W}(\beta + T_j, [\frac{1}{\beta + T_j}(\beta \boldsymbol{W} + \sum_{t:c_{t,j}=1} (\boldsymbol{z}_t - \boldsymbol{\mu}_j)^\top (\boldsymbol{z}_t - \boldsymbol{\mu}_j))]^{-1}) \end{aligned} \tag{5}$$

2.2 Hyper-parameters

For hyper-parameters $\boldsymbol{\lambda}, \boldsymbol{R}$, and $\boldsymbol{W}$ the priors are defined as follows:

$$p(\boldsymbol{\lambda}) \sim \mathcal{N}(\boldsymbol{\mu}_Z, \mathbf{cov}_Z) \quad p(\boldsymbol{R}) \sim \mathcal{W}(1, \mathbf{cov}_Z^{-1}) \quad p(\boldsymbol{W}) \sim \mathcal{W}(1, \mathbf{cov}_Z) \tag{6}$$

variables $\boldsymbol{\mu}_Z \in \mathbb{R}^D$ and $\mathbf{cov}_Z \in \mathbb{R}^{D \times D}$, are respectively the mean and covariance of $\boldsymbol{Z}$. Following the procedure exposed in Sect. 2.1, the posterior distributions on hyper-parameters are obtained straight forward using the mean and precision priors, Eq. 2, as likelihoods in each case:

$$\begin{aligned} p(\boldsymbol{\lambda}|\{\boldsymbol{\mu}_j\}_{j=1}^{k_{\text{rep}}}, \boldsymbol{R}) &\propto \prod_{j=1}^{k_{\text{rep}}} p(\boldsymbol{\mu}_j|\boldsymbol{\lambda}, \boldsymbol{R}) \times p(\boldsymbol{\lambda}) \\ &\sim \mathcal{N}\left((\boldsymbol{\mu}_Z \mathbf{cov}_Z^{-1} + \boldsymbol{R} \sum_{j=1}^{k_{\text{rep}}} \boldsymbol{\mu}_j)(\mathbf{cov}_Z^{-1} + k_{\text{rep}} \boldsymbol{R})^{-1}, (\mathbf{cov}_Z^{-1} + k_{\text{rep}} \boldsymbol{R})^{-1} \right) \end{aligned} \tag{7}$$

$$\begin{aligned} p(\boldsymbol{R}|\{\boldsymbol{\mu}_j\}_{j=1}^{k_{\text{rep}}}, \boldsymbol{\lambda}) &\propto \prod_{j=1}^{k_{\text{rep}}} p(\boldsymbol{\mu}_j|\boldsymbol{\lambda}, \boldsymbol{R}) \times p(\boldsymbol{R}) \\ &\sim \mathcal{W}\left(k_{\text{rep}} + 1, \left[\frac{\mathbf{cov}_Z + \sum_{j=1}^{k_{\text{rep}}} (\boldsymbol{\mu}_j - \boldsymbol{\lambda})^\top (\boldsymbol{\mu}_j - \boldsymbol{\lambda})}{k_{\text{rep}} + 1} \right]^{-1} \right) \end{aligned} \tag{8}$$

$$\begin{aligned} p(\boldsymbol{W}|\{\boldsymbol{S}_j\}_{j=1}^{k_{\text{rep}}}, \beta) &\propto \prod_{j=1}^{k_{\text{rep}}} p(\boldsymbol{S}_j|\beta, \boldsymbol{W}) \times p(\boldsymbol{W}) \\ &\sim \mathcal{W}\left(k_{\text{rep}}\beta + 1, \left[\frac{\mathbf{cov}_Z^{-1} + \sum_{j=1}^{k_{\text{rep}}} \boldsymbol{S}_j}{k_{\text{rep}}\beta + 1} \right]^{-1} \right) \end{aligned} \tag{9}$$

Parameter β remains scalar after conjugacy. According to Rasmussen [16], it has gamma prior of the form:

$$g = \beta - D + 1 \tag{10}$$

$$p(g^{-1}) \sim \mathcal{G}(1, \frac{1}{D}) \quad \rightarrow \quad p(g) \propto g^{-\frac{3}{2}} \exp\{-\frac{D}{2\,g}\} \tag{11}$$

For this parameter the posterior distribution takes the following form:

$$\begin{aligned} p(g|\{\boldsymbol{S}_j\}_{j=1}^{k_{\text{rep}}}, \boldsymbol{W}) &\propto \prod_{j=1}^{k_{\text{rep}}} p(\boldsymbol{S}_j|\beta, \boldsymbol{W}) \times p(g) \\ &\propto (\frac{\beta}{2})^{\frac{k_{\text{rep}}\,\beta\,D}{2}}\, g^{-\frac{3}{2}}\, \Gamma_D(\frac{\beta}{2})^{-k_{\text{rep}}}\, \exp\{-\frac{D}{2\,g}\} \prod_{j=1}^{k_{\text{rep}}} |\boldsymbol{W}\,\boldsymbol{S}_j|^{\frac{\beta}{2}}\, \exp\{-\frac{1}{2}\,\beta\,\mathbf{tr}(\boldsymbol{W}\boldsymbol{S}_j)\} \end{aligned} \tag{12}$$

The later density is not standard form. However, $p(\log(g)|\{\boldsymbol{S}_j\}_{j=1}^{k_{\text{rep}}}, \boldsymbol{W})$ is log-concave, so we may generate independent samples using the Adaptive Rejection Sampling technique (ARS), and transform these samples to get values of β.

2.3 Mixing Proportions and Latent Variables

In this section k is not limited to represented classes. For the mixing proportions π_j, the prior is a symmetric Dirichlet distribution with concentration α/k.

$$p(\{\pi_j\}_{j=1}^k|\alpha) \sim \text{Dir}(\{\alpha/k\}_{j=1}^k) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \prod_{j=1}^k \pi_j^{\alpha/k-1}, \tag{13}$$

where $\Gamma(\cdot)$ is the gamma function. Likewise, the joint distribution for the latent variable $\boldsymbol{c}_t$ has the following form:

$$p(\{c_{t,j}\}_{j=1}^k|\{\pi\}_{j=1}^k) = \prod_{j=1}^k \pi_j^{c_{t,j}}, \quad \{\forall t : \prod_{j=1}^k \pi_j^{T_j}\}, \tag{14}$$

Using the Dirichlet integral type I, the prior is directly written in terms of the latent variable:

$$\begin{aligned} p(\{c_j\}_{j=1}^k|\alpha) &= \int p(\{c_j\}_{j=1}^k|\{\pi_j\}_{j=1}^k)\; p(\{\pi_j\}_{j=1}^k) d\pi_1 \cdots d\pi_k \\ &= \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \prod_{j=1}^k \frac{\Gamma(T_j + \alpha/k)}{\Gamma(\alpha/k)}. \end{aligned} \tag{15}$$

For estimating variable $\boldsymbol{c}_t$, it is required the prior for a single indicator given all others. This is obtained from Eq. 15, keeping all but a single indicator fixed:

$$p(c_{t,j} = 1|\boldsymbol{c}_{-t}, \alpha) = \frac{T_{-t,j} + \alpha/k}{T - 1 + \alpha}. \tag{16}$$

where the subscript $-t$ indicates all the indexes except t and $T_{-t,j}$ is the number of observations, excluding $\boldsymbol{z}_t$, that are associated with component j.
Lastly, an inverse Gamma prior is chosen for parameter α:

$$p(\alpha^{-1}) \sim \mathcal{G}(1,1) \rightarrow p(\alpha) \propto \alpha^{-3/2} \exp\{-1/2\alpha\}. \tag{17}$$

The likelihood for α is derived from Eq. 15, its posterior distribution takes the following form:

$$\begin{aligned} p(\alpha|\{T_j\}_{j=1}^k, k) &= p(\{T_j\}_{j=1}^k|\alpha) \times p(\alpha) \\ &\propto \frac{\alpha^{k-3/2} \exp\{-1/2\alpha\}\Gamma(\alpha)}{\Gamma(T+\alpha)} \end{aligned} \tag{18}$$

Sampling from the later density requires employing ARS. In the limit where $k \rightarrow \infty$, the conditional prior for $\boldsymbol{c}_t$, Eq. 16, becomes:

$$\begin{aligned} \text{Components where } T_{-t,j} > 0: \quad p(c_{t,j} = 1|\boldsymbol{c}_{-i}, \alpha) &= \frac{T_{-i,j}}{T-1+\alpha}, \\ \text{else:} \quad p(\boldsymbol{c}_t \neq \boldsymbol{c}_{t'}, \{\forall t \neq t'\}|\boldsymbol{c}_{-t}, \alpha) &= \frac{\alpha}{T-1+\alpha}. \end{aligned} \tag{19}$$

The posterior is obtained by multiplying the complete likelihood, Eq. 1, and the latent variables prior, Eq. 19:

$$\begin{aligned} \text{Components where } T_{-t,j} > 0: \quad & p(c_{t,j} = 1|\boldsymbol{c}_{-i}, \boldsymbol{\mu}_j, \boldsymbol{S}_j, \alpha) \\ & \propto \frac{T_{-i,j}}{T-1+\alpha} |\boldsymbol{S}_j|^{\frac{1}{2}} \exp\{-\frac{1}{2}(\boldsymbol{z}_t - \boldsymbol{\mu}_j)\, \boldsymbol{S}\, (\boldsymbol{z}_t - \boldsymbol{\mu}_j)^\top\}, \end{aligned} \tag{20}$$

$$\begin{aligned} \text{else:} \quad & p(\boldsymbol{c}_t \neq \boldsymbol{c}_{t'}, \{\forall t \neq t'\}|\boldsymbol{c}_{-t}, \boldsymbol{\lambda}, \boldsymbol{R}, \beta, \boldsymbol{W}, \alpha) \\ & \propto \frac{\alpha}{T-1+\alpha} \int p(\boldsymbol{z}_t|\boldsymbol{\mu}_j, \boldsymbol{S}_j)\, p(\boldsymbol{\mu}_j, \boldsymbol{S}_j|\boldsymbol{\lambda}, \boldsymbol{R}, \beta, \boldsymbol{W})\, d\boldsymbol{\mu}_j\, d\boldsymbol{S}_j. \end{aligned} \tag{21}$$

The likelihood for components with observations other than $\boldsymbol{z}_t$ is Gaussian with parameters $\boldsymbol{\mu}_j$ and $\boldsymbol{S}_j$. On the other hand, for unrepresented classes the likelihood parameters are obtained by sampling from the components priors, as the marginalization of existing parameters is not analytically tractable [16]. When an unrepresented class is chosen, a new class is introduced to the model. Likewise, when a class becomes empty, the class is removed from the model.

3 Experimental Setup

Database. To test our *Infinite Gaussian Fisher Vector* encoding approach (IGFV), we employ both the UCF50 [17] and HMDB51 [11] databases. The

UCF50 database contains realistic videos taken from Youtube, with substantial variation in-camera motion, object appearance, and illumination changes. For concrete testing, we use $N = 5967$ videos concerning 46 human action categories. Following the standard procedure, we perform a leave-one-group-out cross-validation scheme and report the average accuracy over 25 predefined groups [17]. On the other hand, the HMDB51 database is collected from a variety of sources. For the sake of simplicity, we use $N = 6510$ video sequences concerning 51 action categories. Following the proposed protocol, we perform 3-fold cross-validation and report the average accuracy over three predefined train-test splits [11].

Settings. For each video sample, we employ the hand-crafted Improved Dense Trajectory feature estimation technique (iDT), with the code provided by the authors in [20]. Using the default settings, we extract the following trajectory aligned descriptors: Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF), and Motion Boundary Histogram (MBHx, and MBHy). All descriptors are extracted along all valid trajectories and the resulting dimensionality D is 96 for HOG, MBHx, and MBHy, and 108 for HOF.

In practice, using the standard Wishart distribution for sampling model precisions $\boldsymbol{S}_j$, $\boldsymbol{R}$, and $\boldsymbol{W}$ may generate matrices that are not symmetric positive semidefinite (SPD). To avoid this inconvenient, we employ the Frobenius norm positive approximation from [10], that is: For an arbitrary matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$, its nearest SPD Frobenius approximation is set to be $\widehat{\boldsymbol{A}}_F = (\boldsymbol{B} + \boldsymbol{H})/2$, where $\boldsymbol{H}$ is the symmetric polar factor of $\boldsymbol{B} = (\boldsymbol{A} + \boldsymbol{A}^\top)/2$.

We use the ARS algorithm for sampling scalar parameters β and α. In brief, the algorithm employs piecewise exponential functions for approximating any univariate log-concave density $h(x)$ through an envelope (upper hull) and squeezing function (lower hull). Both touch the density function at m sampled points, known as abscissae $(x_1, \ldots, x_m)$. Conventionally, the starting point x_1 is chosen such that $h'(x_1) > 0$, and the final point x_m is chosen such that $h'(x_m) < 0$, where $h'(x) = h(x)/dx$. Even though the method is adaptive and approximated curves will converge to the density function. An erroneous initialization generates ill-posed samples that hamper the proper operation of the algorithm. We solve this issue through a trial-and-error iterative solution. Finally, the samples are obtained as stated in [9].

Training. Initially, we randomly select a subsample of 5000 trajectories per category from the training set. Then, using PCA we select the most relevant attributes until 90% of input variability is preserved. Later, we employ the spatio-temporal pyramid technique for distributing the training partition into cells. For each spatio-temporal cell, we estimate an IGMM codebook using the procedure exposed in Sect. 2. The model starts with a single component, then 1000 iterations of Gibbs sampling are performed for updating all parameters and hyper-parameters iteratively from their posterior distribution, with 800 "burn in iterations". From the remaining 200 repetitions, we use Bayesian Information Criterion (BIC) for choosing the best available mixture model. Afterward,

the conventional FV encoding technique is employed for representing locally described samples as super-vectors [7]. In short, the method quantifies the similarity between a video sample and trained IGMM model, named codebook. To the resulting super-vector, we apply a Power Normalization (PN) followed by the L2-Normalization. The above procedure is performed per descriptor. Ultimately, all four IGFV representations are concatenated together.

For the classification step, we use a one-vs-all Linear SVM with regularization parameter equal to 100. Figure 1 summarizes the IGFV training pipeline. It is worth noting that the feature extraction was performed in C++ and the remaining experiments in MATLAB.

4 Results and Discussions

ARS Correction. Figure 2 shows the proposed correction for ARS initialization, when sampling parameter α. In Fig. 2a, we see that abscissae x_1 and x_3 are chosen according to the conventional criteria (i.e., $h'(x_1) > 0$ and $h'(x_3) < 0$). Though $x_3 = 2.38$ satisfy the restriction, $m_3 = -0.002$. This value creates wide upper hull that may generate ill-posed samples. In this case, $\widehat{\alpha} = 3019$ when the abscissae range (most probable values) is around $[1.5, 2.4]$. To solve this problem, we iteratively increase x_3 by 50% until the upper hull is relatively constrained. Figure 2b is obtained through this procedure. Here, $x_3 = 3.6$, $m_3 = -2.18$, and the sampled parameter is $\widehat{\alpha} = 2.05$.

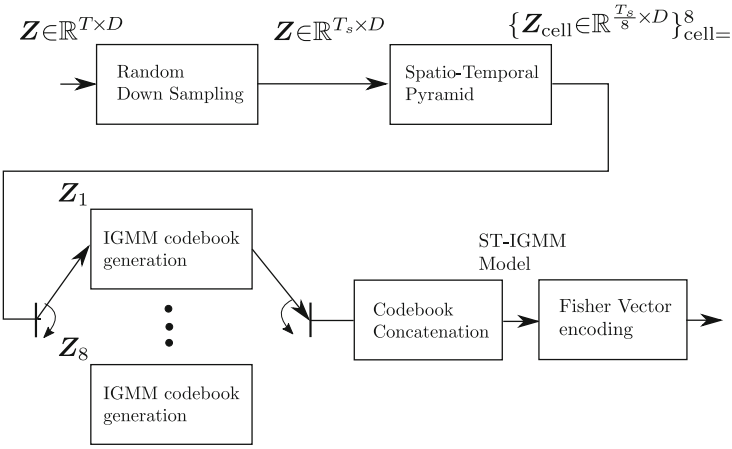
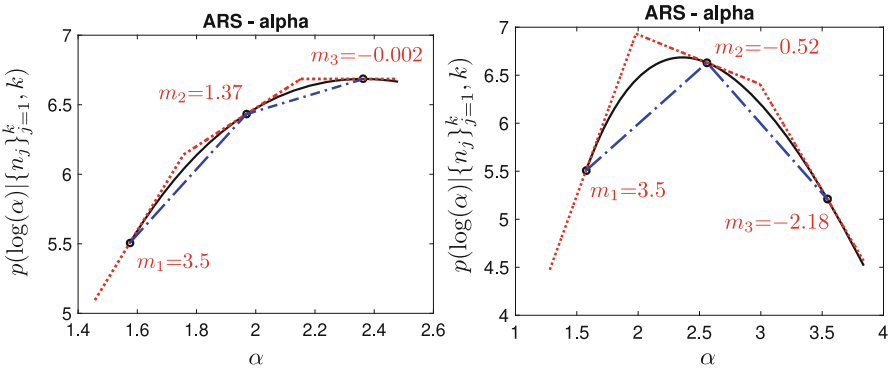


Fig. 1. Sketch of the proposed IGFV data encoding technique.

Confusion Matrices. Figure 3 shows the obtained confusion matrices using linear SVM for both employed databases. The proposal achieves $88.5 \pm 4.07\%$ and 57.6 ± 1.93 of mean accuracy, within the cross-validation scheme for each



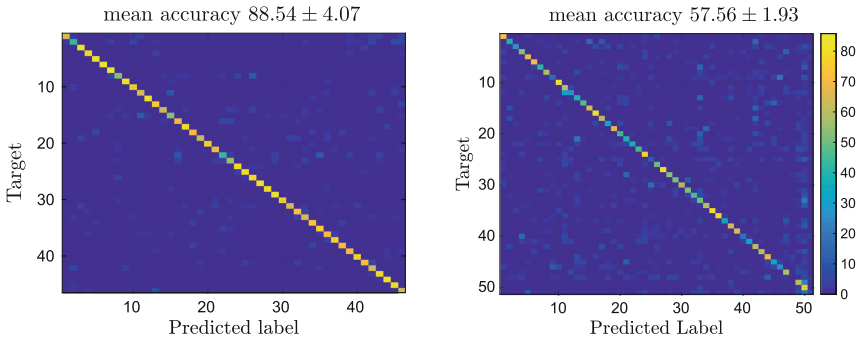
(a) Conventional initialization ($\widehat{\alpha}$=3019) (b) Corrected initialization ($\widehat{\alpha}$=2.0546)

Fig. 2. Adaptive Rejection Sampling with initialization correction. ——concave log-density, - · - lower hull, and · · · upper hull. Circular points indicate the initial three abscissae (x_1, x_2, x_3).

database. From a visual inspection on Fig. 3a, the system demonstrates the ability to discriminate among human actions, with slight errors in a few categories. On the other hand, the visual inspection on Fig. 3b shows how difficult it is for the system to classify actions from the HMDB51 dataset. When reviewed in detail, the provided bounding boxes from some videos do not correspond, partially or entirely, to the reported activity. Thus, the bad performance of the system may be explained by this issue, considering the importance of an effective human detection within iDT feature estimation.

Comparison with the State of the Art. In turn, Table 1 presents a comparative study among similar feature encoding approaches for human action recognition. In this study, we are analyzing properties and comparing characteristics among encoding methodologies. Thus, for feature extraction and classification we standard methods for the sake of comparison. Employed benchmarks have in common the following considerations: (i) are tested in both UCF50 and HMDB51 databases, (ii) employ the iDT feature estimation technique, (iii) perform classification through linear SVM. In particular, ST-VLAD [5] and SFV-STP [20], follow all requirements. However, they require extensive cross-validation for estimating the number of components k in their codebook. Moreover, ST-VLAD also needs searching the number of Spatio-temporal groups, which for both SFV-STP [20] and our IGFV are fixed to 8 cells, one division for each spatial and temporal axis.

Our main contribution is the automation of a methodology that conventionally requires extensive cross-validation. Thus, the slight drop in accuracy from our method, when compared to benchmarks, is compensated with computational savings, because it both reduces number of operations and memory requirements.



(a) UCF50 (b) HMDB51

Fig. 3. Confusion matrices from Human Action Recognition in UCF50 and HMDB51 databases.

In particular, conventional GMM codebooks perform maximum likelihood estimation, through EM, of model parameters. Such optimization, performs a large number of iterations (in our case 500), and its convergence depends on initialization. Furthermore, this method needs fixing the number of components k before optimizing all other parameters (means, precision, and priors). Authors in [20], proposed searching k in a set comprising 10 different values and then selecting the best value according to classification performance. This approach requires cross-validating k in an operation that increases the number of iterations by 10 times the number of folds in the cross-validation. In terms of memory requirements, it requires storing all parameters 10 times, until the best k is chosen.

The drop in accuracy from our method could be attributed to the precisions sampled by the IGMM. In our case, IGMM samples complete precision matrices that are considering the correlation between attributes. For us, this is a disadvantage as the IGMM codebook suffers from low resolution, i.e., complete Gaussians can explain massive data clusters. Meanwhile, benchmarks approaches constrain covariances to diagonal or spherical matrices. Thus, the estimated number of components from IGMM is in the order of tens, whereas the exhaustive search from benchmarks converges to hundreds of components. This is an interesting result that demonstrates the quality of IGMM estimated codebooks, as fewer components allows the codification of discriminant visual information.

Table 1. Comparison with similar approaches on UCF50 and HMDB51 datasets.

Methods	Components	UCF50 [%]	HMDB51 [%]
ST-VLAD [5]	$k=256$ (exhaustive search)	90.7	59.0
SFV+STP [20]	$k=256$ (exhaustive search)	91.7	60.1
IGFV (proposal)	Automatic	88.5	57.6

5 Conclusions

We introduced a novel Infinite Gaussian Fisher Vector feature encoding framework to support video-based Human Action Recognition (IGFV). Our approach is fully automatic, allowing every parameter in the model to be updated hierarchically through the MCMC algorithm Gibbs sampling. The IGFV encoding allows revealing a set of discriminant local spatio-temporal features for enabling the precise codification of visual information, with competitive recognition results and computational savings. In particular, the infinite limit on the number of Gaussian components evades estimating this parameter through extensive cross-validation, which drastically reduces the number of operations and memory requirements for performing HAR. Attained results on both UCF50 and HMDB51 database showed that our proposal correctly classified 88.5% and 57.6% of human actions under the specific cross-validation of each dataset. Our IGFV obtained promising results that are comparable with state-of-art encoding approaches. Furthermore, it outperforms those approaches considering the trade-off between accuracy and computational complexity, as our proposal reduces both number of operations and memory requirements.

As future work, authors will evaluate alternatives for enhancing the resolution from IGMM codebooks, such as placing diagonal or spherical constrains to the sampled precision. We are convinced that all the already mentioned benefits from Bayesian inference and Dirichlet processes, combined with an enhanced model resolution, will yield better performance in human action recognition tasks.

Acknowledgments. Under grants provided by the project: "Prototipo de un sistema de recuperación de información por contenido orientado a la localización y clasificación de grupos de microcalcificaciones en mamografías - PROTOCAM", CV E6-19-1, from the VIIE-UTP. Also, J. Fernández is partially funded by the Colciencias program: *Jóvenes investigadores e innovadores-Convocatoria 812 de 2018*, and by the project "Sitema de clasificación de videos basado en técnicas de representación utilizando métodos núcleo e inferencia bayesiana", CV E6-19-2, from the VIIE-UTP.

References

1. Bloom, V., Argyriou, V., Makris, D.: Linear latent low dimensional space for online early action recognition and prediction. Pattern Recogn. **72**, 532–547 (2017)
2. Borges, P.V.K., Conci, N., Cavallaro, A.: Video-based human behavior understanding: a survey. IEEE Trans. Circuits Syst. Video Technol. **23**(11), 1993–2008 (2013)
3. Carmona, J., Climent, J.: Human action recognition by means of subtensor projections and dense trajectories. Pattern Recogn. **81**, 443–455 (2018)
4. Chen, T., Morris, J., Martin, E.: Probability density estimation via an infinite Gaussian mixture model: application to statistical process monitoring. J. R. Stat. Soc. Ser. C Appl. Stat. **55**(5), 699–715 (2006)
5. Duta, I.C., Ionescu, B., Aizawa, K., Sebe, N.: Spatio-temporal VLAD encoding for human action recognition in videos. In: Amsaleg, L., Guðmundsson, G., Gurrin, C., Jónsson, B., Satoh, S. (eds.) MMM 2017. LNCS, vol. 10132, pp. 365–378. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-51811-4_30

6. Fan, W., Bouguila, N., Liu, X.: A nonparametric Bayesian learning model using accelerated variational inference and feature selection. Pattern Anal. Appl. **22**(1), 63–74 (2019)
7. Fernández-Ramírez, J., Álvarez-Meza, A., Orozco-Gutiérrez, Á.: Video-based human action recognition using kernel relevance analysis. In: Bebis, G., et al. (eds.) ISVC 2018. LNCS, vol. 11241, pp. 116–125. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-03801-4_11
8. Field, M., Stirling, D., Pan, Z., Ros, M., Naghdy, F.: Recognizing human motions through mixture modeling of inertial data. Pattern Recogn. **48**(8), 2394–2406 (2015)
9. Gilks, W.R., Wild, P.: Adaptive rejection sampling for Gibbs sampling. J. R. Stat. Soc. Ser. C (Appl. Stat.) **41**(2), 337–348 (1992)
10. Higham, N.: Computing a nearest symmetric positive semidefinite matrix. Linear Algebra Appl. **103**(C), 103–118 (1988)
11. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: Proceedings of the International Conference on Computer Vision (ICCV) (2011)
12. Li, Q., Cheng, H., Zhou, Y., Huo, G.: Human action recognition using improved salient dense trajectories. Comput. Intell. Neurosci. **2016**, 1–11 (2016)
13. Ma, C.Y., Chen, M.H., Kira, Z., AlRegib, G.: TS-LSTM and temporal-inception: exploiting spatiotemporal dynamics for activity recognition. Sig. Process. Image Commun. **71**, 76–87 (2019)
14. Priya, T., Prasad, S., Wu, H.: Superpixels for spatially reinforced Bayesian classification of hyperspectral images. IEEE Geosci. Remote Sens. Lett. **12**(5), 1071–1075 (2015)
15. Qian, Y., Sengupta, B.: Pillar networks: combining parametric with non-parametric methods for action recognition. Robot. Autonomous Syst. **118**, 47–54 (2019)
16. Rasmussen, C.: The infinite Gaussian mixture model, pp. 554–559 (2000)
17. Reddy, K.K., Shah, M.: Recognizing 50 human action categories of web videos. Mach. Vis. Appl. **24**(5), 971–981 (2013)
18. Sicre, R., Nicolas, H.: Improved Gaussian mixture model for the task of object tracking. In: Real, P., Diaz-Pernil, D., Molina-Abril, H., Berciano, A., Kropatsch, W. (eds.) CAIP 2011. LNCS, vol. 6855, pp. 389–396. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23678-5_46
19. Uijlings, J., Duta, I.C., Sangineto, E., Sebe, N.: Video classification with densely extracted HOG/HOF/MBH features: an evaluation of the accuracy/computational efficiency trade-off. Int. J. Multimedia Inf. Retrieval **4**(1), 33–44 (2015)
20. Wang, H., Oneata, D., Verbeek, J., Schmid, C.: A robust and efficient video representation for action recognition. Int. J. Comput. Vis. **119**(3), 219–238 (2016)
21. Wang, S., Hou, Y., Li, Z., Dong, J., Tang, C.: Combining convnets with hand-crafted features for action recognition based on an HMM-SVM classifier. Multimedia Tools Appl. **77**(15), 18983–18998 (2018)
22. Weng, J., Weng, C., Yuan, J., Liu, Z.: Discriminative spatio-temporal pattern discovery for 3D action recognition. IEEE Trans. Circuits Syst. Video Technol. **29**(4), 1077–1089 (2019)