2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

One-shot learning of temporal sequences using a distance dependent Chinese Restaurant Process

Carlos Orrite
Aragon Insititute of Engineering Research
University of Zaragoza, Spain
Email: corrite@unizar.es

Mario Rodriguez
Aragon Insititute of Engineering Research
University of Zaragoza, Spain
Email: mrodrigo@unizar.es

Carlos Medrano
EduQTech, E.U. Politecnica
University of Zaragoza, Spain
Email: ctmedra@unizar.es

***Abstract*—Activity recognition in videos is a challenging task, mainly if a scarce number of samples is available for modelling the problem. The task becomes even harder when using generative models such as mixture models or Hidden Markov Models (HMMs), as they demand a lot of samples to determinate their parameters. Additionally, these models rely on the appropriate selection of some parameters, for instance the number of hidden states. Therefore, we propose in this paper the creation of a Universal Background Model (UBM) of features, using videos from public datasets, applied to the activity encoding and an unsupervised modelling of the activities with a distance dependent Chinese Restaurant Process (ddCRP), where the number of states is automatically determined by the process.**

In order to classify an incoming video-sequence we propose to model it as a ddCRP distribution and to apply a nearest neighbour algorithm based on a kernel between distributions. To carry out this process we use a Probability Product Kernel (PPK) algorithm by previously mapping the ddCRP into a HMM with discrete observations. Preliminary experiments in two public data sets, as Weizmann and KTH, show that this proposal achieves state-of-the-art results.

I. Introduction

Understanding and interpreting human behaviours based on video analysis is an excellent tool to improve human-machine interaction and therefore it has witnessed tremendous progress in the last years. Some applications in video-surveillance, gaming or Ambient Assisted Living need the recognition of the activities using fixed viewpoint cameras and therefore the clutter and variabilities produced by background change or moving viewpoint are significantly reduced.

However, after the installation, the system should be re-trained again as any previously collected sequences may not be representative of the new environment. Although the performance is constrained by the number of labelled sequences used for training, collecting and labelling large amount of data for the particular scenario is infeasible.

Moreover, in some video-surveillance applications, the system should be able to detect unusual events and trigger some kind of alarm. These unusual events are difficult to obtain due to their rare frequency and even more to obtain enough examples for a trustworthy training of the activity.

Therefore, we propose in this paper the creation of a Universal Background Model (UBM) of features, using videos from public datasets, applied to the activity encoding. In this way, a Gaussian Mixture Model (GMM) is created from these videos and afterwards, a soft-assignment of the one-shot can be carried out to create the observation streaming.

Dirichlet process (DP) mixture models provide a valuable suite of flexible clustering algorithms. DP mixtures can be described via the Chinese restaurant process (CRP), which is fancifully described by a sequence of customers sitting down at the tables of a Chinese restaurant. Originally, each customer sits at a previously occupied table with probability proportional to the number of customers already sitting there, and at a new table with probability related to a concentration parameter. Rencently, a new approach denoted Distance Dependent Chinese Restaurant Process (ddCRP) has emerged [1]. The ddCRP is a flexible class of distributions over partitions that allows for dependencies among elements. This class can be used to model many kinds of dependencies among data samples in infinite clustering models, including dependencies arising from time, space, and network connectivity. In our approach we will consider the time relationship among the customers arriving to the restaurant which will be helpful to model the temporal dependency in video-sequences.

The central computational problem for ddCRP modelling is posterior inference, determining the conditional distribution of the latent variables given the observations. Regardless of the likelihood model, the posterior will be intractable to compute. Recently, a fast approach for inference in Dirichlet process mixture models has been proposed [2]. In that paper authors propose a sequential updating and greedy search algorithm. We will modify this algorithm in order to infer the latent variables in the ddCRP.

Once the action sequence has been modelled by a ddCRP, next step consists in classifying a new video sequence to a specific activity class learned from just one sample per action. In [2] authors use the pseudo-marginal likelihood, defined as the product of conditional predictive ordinates to adress this issue. However, this approach is very time consuming for large sequences. Additionaly, recent developments have shown the better performance of kernels for classification even in the case of comparing distributions, as it happens in our approach. In [3] authors considered kernels between HMMs using the Probability Product Kernel (PPK) to compute the affinity between distributions. Therefore, our approach consists in mapping data points, in the input space, into distributions over the sample space and a general inner product is then

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

evaluated as the integral of the product of pairs of distributions. To compute the kernel for two sequences we train a ddCRP for each one, as described in Section III, and then compute the kernel using the learned ddCRP prior probability of a table assignment and the conditional likelihood of the observation.

The rest of the paper is divided as follows. Section II explains how to obtain the sequence of observations using a UBM. Section III explains how to get a ddCRP from a sequence of observations. Section IV deals with the comparison of two distributions by the PPK in order to recognize an unknown temporal sequence. Section V presents our experimental validation and section VI states some conclusions and proposals for future work.

II. Video Encoding

The lack of sufficient data in a One-shot learning scenario forces to use source videos, widely available, to perform the video encoding. Firstly, a UBM is modelled by a GMM using the source videos and afterwards, a soft-assignment process is carried out to create an observation. This process is repeated in temporal windows. Next, we describe the temporal activity modelling by a GMM following a soft-assignment to a Bag of Features (BoF) approach. Later, we explain the motivation for using a sliding window to code temporal information given as a result a sequence of observations to train a Dirichlet process. Figure 1 shows the process where each feature from a windowed video is soft-assigned to its corresponding BoF.

A. Soft-assignment-BoF

The proposed encoding uses the Imprved Dense Trajectories (IDT)[4] features extracted from the activity videos and models the features space through a GMM. The number of clusters M can vary a lot in different approaches and empirically has been proven that a large number, in the order of thousands, is appropriate for BoF encoding.

From each video activity, a set of IDT feature vectors $\mathbf{Q} = \{\mathbf{q}_j\}$, $\mathbf{q}_j \in \mathbb{R}^D$ is extracted. Using the feature vectors of the training examples a GMM, $\lambda = \{\omega_i, \mu_i, \Sigma_i\}$, is calculated. The general model of GMM supports full covariance matrices, but diagonal covariance matrices can satisfactorily approximate the original density modelling with a higher order GMM and they are computationally more efficient. Therefore, the framework uses diagonal covariance matrices and in addition it disregards GMM weights obtaining a simplified model $\lambda = \{\mu_i, \Sigma_i\}$.

Using the simplified GMM, $\lambda = \{\mu_i, \Sigma_i\}$, the activity in a video is encoded with a BoF where each bin value v_{λ_i} is calculated by proportionally adding the contributions of every extracted feature $\mathbf{q}_j$ to the specific Gaussian λ_i, as expressed in Equation 1.

$$v_{\lambda_i} = \frac{1}{L}\sum_{j=1}^{L} p(\lambda_i|\mathbf{q}_j) = \frac{1}{L}\sum_{j=1}^{L} \frac{\mathcal{N}(\mathbf{q}_j;\mu_i,\Sigma_i)}{\sum_{m=1}^{M}\mathcal{N}(\mathbf{q}_j;\mu_m,\Sigma_m)} \quad (1)$$

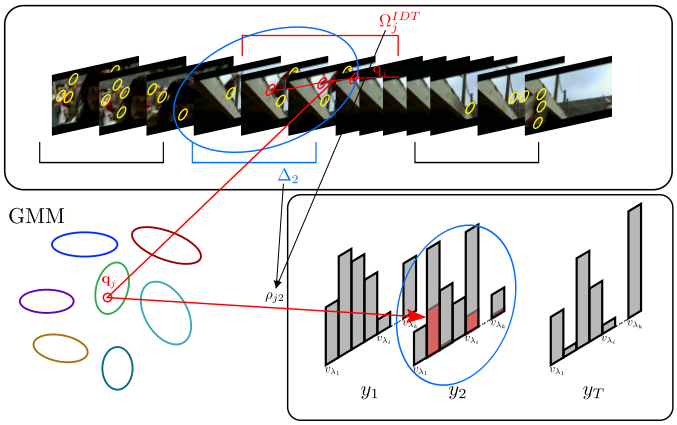

Fig. 1. Feature extraction, $\mathbf{q}_j$ and contribution to a BoF sequence of Windowed video (red) using a GMM for feature space modelling.

where L is the number of features extracted, and M is the number of Gaussians in the GMM as well as the number of bins of BoF histogram.

B. Temporal Sliding Window

The former video encoding loses the long-term temporal information of the activity if performed along the whole video. Therefore, a temporally windowed Soft-assignment-BoF encoding is proposed. Using a stream of N_Δ frames sliding windows, the long-term temporal information is kept in a sequence of BoF, $\mathcal{Y} = \{y_1, \cdots, y_T\}$. The encoding uses IDT as descriptors which are computed through a temporal window of length N_Ω, generally different to N_Δ.

Each IDT, $\mathbf{q}_j$, is extracted from a frame window Ω_j^{IDT}. On the other hand, the video is divided into sliding frame windows Δ_t. The number of frames and the position of the windows are generally different, and each IDT influences the each BoF y_t proportionally given the Equation 2

$$\rho_{jt} = \frac{|\Delta_t \cap \Omega_j^{IDT}|}{N_\Delta} \quad (2)$$

Each bin value, $v_{\lambda_i}^t$, associates to a specific BoF, y_t, is then calculated using Equation 3

$$v_{\lambda_i}^t = \frac{1}{\sum_{j=1}^{L}\rho_{jt}}\sum_{j=1}^{L}\frac{\rho_{jt}\mathcal{N}(\mathbf{q}_j;\mu_i^t,\Sigma_i)}{\sum_{m=1}^{M}\mathcal{N}(\mathbf{q}_j;\mu_m^t,\Sigma_m)} \quad (3)$$

III. Dirichlet Process

To better understand the ddCRP it is helpful to compare it with the Chinese Restaurant Process. So, first we will describe the traditional CRP, later we will introduce the ddCRP and finally, we will pay attention to the central computational problem for this approach, i.e., modelling its posterior inference.

A. Chinese Restaurant Process (CRP)

In the traditional CRP, the probability of a customer sitting at a table is computed from the number of other customers already sitting at that table. Let z_i denote the table assignment

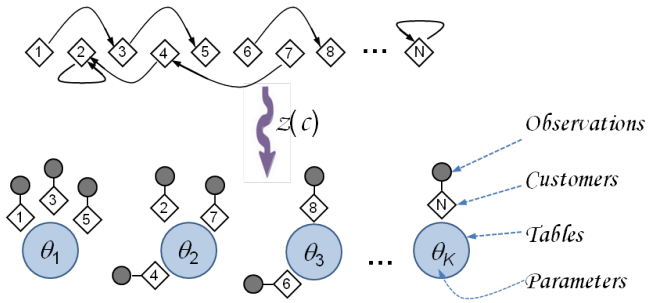

Fig. 2. Distance dependent Chinese Restaurant Process.

of the ith customer, let n_k denote the number of customers sitting at table k and K the number of occupied tables. The traditional CRP seats newest ith customer to the table z_i according to the following probability distribution

$$p(z_i = k \mid z_{-i}, \alpha) \propto \begin{cases} \frac{n_k}{\alpha+i} & \text{for} \quad k \leq K \\ \frac{\alpha}{\alpha+i} & \text{for} \quad k = K+1 \end{cases} \tag{4}$$

where α is a given scaling parameter. When all N customers have been seated, their table assignments provide a random partition which is invariant to the order they sat down.

B. Distance dependent CRP (ddCRP)

The ddCRP was introduced as a flexible class of distributions over partitions that allows for dependencies between the elements [1], see Figure 2. These dependencies may arise from time, space or network connectivity. Continuing with the culinary metaphor, the newest ith customer chooses to sit down with some other customer j (denoted as $c_i = j$) with a probability proportional to a decreasing function of the distance between the two: $f(d_{ij})$ or by himself with a probability proportional to α

$$p(c_i = j \mid D, \alpha) \propto \begin{cases} f(d_{ij}) & \text{if} \quad i \neq j \\ \alpha & \text{if} \quad i = j \end{cases} \tag{5}$$

where f stands for a decay function and D denotes the set of all distance measurements between customers i and j. In general, the decay function mediates on how distances between the customers affect the resulting distribution over partitions.

C. Posterior inference

For ddCRP models, the state of the chain is defined by c_i, the customer assignments for each data point. Let $\eta = \{D, \alpha, f, G\}$ denoting the set of model hypermarameters. It contains the distances D, the scaling factor α, the decay function f, and the base measure G. Let y denote the observations. The probability of each customer i to sit down with the customer j in the same table is given by:

$$p(c_i = j \mid c_{-i}, y, \eta) \propto p(c_i = j \mid D, f, \alpha) \cdot p(y \mid c_{-i}, c_i = j, G) \tag{6}$$

The first term is the ddCRP prior. The second term is the likelihood of the observations under the partition given by previous table assignments (c_{-i}) and the new assignment ($c_i = j$).

Regardless of the likelihood model, the posterior will be intractable to compute. Recently, a fast approach for posterior inference in Dirichlet process mixture models has been proposed [2]. In that paper authors proposed to factorize the Dirichlet process as a product of a prior on the partition of subjects into tables and independent priors on the parameters within each table. Adding subjects one at a time, they allocate subjects to the table that maximizes the conditional posterior probability given their data and the allocation of previous subjects, while also updating the posterior distribution of the table-specific parameters. Next section explains in detail how we follow this approach for the recognition of temporal sequences.

D. Inferring the latent variables

The input of our model is a set of D video sequences of different lengths. Each video sequence d is encoded in N_d observations indexed by n. So, the input of the system is a set of temporal templates Y which are the BoF sequences encoded from each video. The output is a sequence of table assignment, i.e., hidden states, $\mathbf{Z_d} = \{z_1, \ldots, z_n, \ldots, z_{N_d}\}$. The total number K of tables (states) is estimated as part of the inference process.

For the sake of clarity, let us describe a representation using the stick-breaking formalism. The parameters in this representation have the following distributions:

$$\pi_k \sim ddCRP(\alpha, f, D) \tag{7}$$

$$z_k \sim \pi_k \tag{8}$$

$$\theta_k \sim G \tag{9}$$

$$y_k \sim F(\theta_k) \tag{10}$$

We accomplish the inference of the latent variables by means of online inference. The process starts assigning the first customer to the first table, i.e., $z_1 = 1$. Afterwards, the new customer assignment can be carried out in two different ways: either customer i implicitly creates a new table with his seating choice, or he connects to an existing table. The table assignment k that maximizes the posterior probability is:

$$\hat{k} = arg\max_k \{p(z_i = k \mid z_{-i}, y_i, y_{-i})\} \tag{11}$$

where the Dirichlet hyperparameters have been omitted from the notation for simplicity.

Using Bayes rule and dropping variables due to conditional independence, this posterior can be decomposed into two terms: the current prior over table assignments and the likelihood of the observed temporal template.

$$\hat{k} = arg\max_k \{\underbrace{p(z_i = k \mid z_{-i})}_{\pi_{ik}} \cdot \underbrace{p(y_i \mid y_{-i}, z_{-i}, z_i = k)}_{\mathcal{L}_k(y_i)}\} \tag{12}$$

The first term π_{ik} is the prior probability of a table assignment and is modelled by a distance dependent CRP, as described in (5). We are assigning customers to tables depending on the people previously sitting there. The new customer decide to sit on an already busy table k with probability proportional to the temporal distance to that customer who occupied that table first.

$$\pi_{ik} \propto \begin{cases} f(max\{d_{ij}\})\forall z_j = k & \text{for} \quad k \leq K \\ \alpha & \text{for} \quad k = K+1 \end{cases} \tag{13}$$

where d_{ij} measures the time difference between costumers i and j and f is a decay function. In our experiments we have considered the exponential decay fucntion $f(d) = e^{-d/a}$, as suggested in [1], where a is a parameter. There exists a relationship between parameters a and α and the total number K of occupied tables. So, when a increases K decreases and on the contrary, when α increases K increases as well. We can fix parameter a and let the concentration parameter α controls the degree of occupancy.

The second term $\mathcal{L}_k(y_i)$ in (12) is the conditional likelihood of y_i given allocation to table k and the table allocation for customers $1, \ldots, i-1$, and must be computed by integrating over the prior conditional Dirichlet distribution G over the probability distribution $\theta_k = \{\theta_{k1}, \cdots, \theta_{kM}\}$.

$$\mathcal{L}_k(y_i) = \int_{\theta_k} p(y_i|\theta_k) \cdot G(\theta_k|y_{-i}, z_{-i}, z_i = k) d\theta_k \tag{14}$$

The first term inside the integral of equation (14) is the multinomial distribution $p(y_i|\theta_k) = \prod_m^M p(\theta_{km})^{y_{i,m}}$ and it describes the likelihood of multiple draws from the discrete distribution θ_k of the k-th table. The second term inside the integral is the conditional Dirichlet distribution G, which is a product of the multinomial distribution of all past observations and a base Dirichlet distribution.

The final form of the likelihood $\mathcal{L}$ conveniently reduces to a product of simple fractions [5],

$$\mathcal{L} \propto \prod_m^M \left[\frac{h(k,m) + \beta_0/M}{\sum_{m'} h(k,m') + \beta_0} \right]^{y_{i,m}} \tag{15}$$

where $h(k,m)$ is the total count of flow vectors in bin m accumulated from all motion histograms that belong to table k and $y_{i,m}$ is the count of the m-th bin in the motion histogram y_i. The hyper-parameter β_0 is set to 1 for all experiments.

Since this inference algorithm only adds new unoccupied tables, i.e., consider a new state added to the previous ones and never removes past states, the order of states is always preserved. This property allows us to accumulate the transition matrix in a single pass by keeping a record of the counts over tables (states) for each video footage.

E. Allowing parameter α to be unknown

As we have mentioned before, the role of the concentration parameter α in relation to the allocation of customers to tables. In order to allow unknown α, we modelled π_{ik} with a mixture of DPs with R different concentration parameters $\alpha = (\alpha_1, \cdots, \alpha_R)'$.

Naming $\phi_{-i}(\alpha_r) = p(\alpha = \alpha_r|z_{-i}, y_{-i})$ and $\pi_{ikr} = p(z_i = k|\alpha = \alpha_r, z_{-i}, y_{-i})$, we obtain the following modification to (12):

$$\hat{k} = arg\max_k \left\{ \sum_{r=1}^{R} \phi_{-i}(\alpha_r) \pi_{ikr} \mathcal{L}_k(y_i) \right\}, k = 1, \cdots, K+1, \tag{16}$$

which is obtained by marginalizing over the posterior for α given the histograms y_{-i} and allocation z_{-i} for customers $1, ...i-1$. Finally, we have the following updated probabilities

$$\phi_i(\alpha_r) = p(\alpha = \alpha_r|z_i, y_i) = \frac{\phi_{-i}(\alpha_r)\pi_{z_i r}}{\sum_s \phi_{-i}(\alpha_s)\pi_{z_i s}} \tag{17}$$

F. Sequence classification

For the task of human motion classification, the goal is to classify a new video sequence to a specific activity class learned from just one example. So, the activity categorization is determined by the aspect corresponding to the highest $P(\theta_w|\mathbf{y})$, that is activity class $w^* = arg\max_w P(\theta_w|\mathbf{y})$.

Using Bayes $P(\theta_w|\mathbf{y}) = P(\mathbf{y}|\theta_w) \cdot P(\theta_w)/P(\mathbf{y})$. As the denominator does not depend on the class w, the maximization takes place considering just the numerator. In order to compute it we use the pseudo-marginal likelihood, which is defined as the product of conditional predictive ordinates [2] as follows,

$$\begin{aligned} p(\mathbf{y}|\theta) &= \prod_{t=1}^{T} p(y_t|y_{-t}, z_{-t}) \\ &= \prod_{t=1}^{T} \int_\theta p(y_t|\theta) \cdot G(\theta|y_{-t}, z_{-t}) d\theta \\ &= \prod_{t=1}^{T} \sum_{k=1}^{K} \pi_{tk} \int_{\theta_k} p(y_t|\theta_k) \cdot G(\theta_k|y_{-t}, z_{-t}) d\theta_k \\ &= \prod_{t=1}^{T} \sum_{k=1}^{K} \pi_{tk} \cdot \mathcal{L}_k(y_t) \end{aligned} \tag{18}$$

where the class w has been omitted for the sake of clarity.

As the length of every sequence can be very long, the computation of the the pseudo-marginal likelihood can be very time consuming. In addition, recent developments [3] [6] have shown the better performance of kernels for classification, even in the case of comparing distributions as it happens in our approach.

IV. DDCRP AND KERNELS

The new approach consists in mapping data points in the input space into distributions over the sample space and a general inner product is then evaluated as the integral of the product of pairs of distributions. To compute the kernel for two sequences χ and χ' of lengths T_χ and $T_{\chi'}$, we train a ddCRP for each sequence, as explained in section III, and then we

compute the kernel using the learned ddCRP prior probabiliy of a table asingment π and the conditional likelihood of the observation $\mathcal{L}$ for a user-specified sequence length T as done in [6], where T can be chosen according to some heuristic.

More specifically, for $p(\mathbf{y}|\theta)$ we consider a first-order HMM with discrete emissions. The probability model of sequence $p(\mathbf{y}|\theta)$ where $y = \{y_1, \ldots, y_T\}$ is a sequence of length T where each observation vector is $y_t \in \Re^d$. The HMM has a hidden state (or table assignation in the ddCRP) at each time point $z = \{z_1, \ldots, z_T\}$ where each state (table) takes a discrete value $z_t = 1, \ldots, K$. The likelihood of the HMM factorizes as follows:

$$p(\mathbf{y}|\theta) = \sum_{z_0,\ldots,z_T} p(y_0|z_0) \prod_{t=1}^{T} p(y_t|z_t)p(z_t|z_{t-1}) \tag{19}$$

where $p(z_t|z_{t-1})$ is the transition matrix computed from the ddCRP in a single pass by keeping a record of the counts over tables (states) for each video footage. $p(y_t|z_t)$ is the emission distribution probability given by the join of histograms assigned to each table.

A. Probability product kernels

A natural choice of kernel between HMMs is the PPK described in [3] since it computes an affinity between distributions. The kernel is computed in closed form for latent distributions such HMMs and can be appropriated for ddCRPs. In this way, the generalized inner product is found by integrating a product of the distributions of pairs of data sequences over the space of all potential observable sequences $\chi : \kappa(p(\mathbf{y} \mid \theta), p(\mathbf{y} \mid \theta')) = \int p(\mathbf{y} \mid \theta)^\rho p(\mathbf{y} \mid \theta')^\rho d\mathbf{y}$. When $\rho = 1/2$, the PPK becomes the classic Bhattacharyya affinity metric between two probability distributions. The Bhattacharyya affinity is favoured over other probabilistic divergences and affinities such as Kullback-Leibler (KL) divergence because it is computable in positive semi-definite (it is a Mercer kernel).

Following [6] we use the factorization of the HMM to compute the $\kappa(\theta, \theta')$ as follows:

$$\begin{aligned}
\kappa(\theta, \theta') &= \int p(\mathbf{y} \mid \theta)^\rho p(\mathbf{y} \mid \theta')^\rho d\mathbf{y} \\
&= \sum_{y_0 \ldots y_T} \sum_{z_0 \ldots z_T} \prod_{t=0}^{T} p(y_t|z_t)^\rho p(z_t|z_{t-1})^\rho \\
&\quad \sum_{z'_0 \ldots z'_T} \prod_{t=0}^{T} p'(y_t|z'_t)^\rho p'(z'_t|z'_{t-1})^\rho \\
&= \sum_{z_T} \sum_{z'_T} \Psi(z_T, z'_T) \prod_{t=1}^{T} \sum_{z_{t-1}} \sum_{z'_{t-1}} p(z_t|z_{t-1})^\rho p(z'_t|z'_{t-1})^\rho \\
&\quad \Psi(z_{t-1}, z'_{t-1}) p(z_0)^\rho p'(z'_0)^\rho
\end{aligned} \tag{20}$$

where $\Psi(z_t, z'_t)$ stands for an elementary kernel computed as:

$$\Psi(z_t, z'_t) = \sum_{y_t} p(y_t|z_t)^\rho p'(y_t|z'_t)^\rho \tag{21}$$

This factorization has been used in a previous paper [6] to provide an efficient iterative method to calculate the PKK for Gaussians distributions. We adapt this formulation to take into account discrete emissions, dealing with ddCRPs having different number of tables too.

We also follow a standard normalization of the kernel, a typical pre-processing step used in the literature.

$$\hat{\kappa}(\theta, \theta') \leftarrow \frac{\hat{\kappa}(\theta, \theta')}{\sqrt{\hat{\kappa}(\theta, \theta)}\sqrt{\hat{\kappa}(\theta', \theta')}} \tag{22}$$

V. Experiments and Results

A. Datasets

The proposed algorithm is trained using human motion information from external video sources, as described in Section II-A. Our method is evaluated using several datasets that accomplish the source and target domain constraints. We have selected three source domain datasets that include a high variability in unconstrained video clips that simulate the easily obtainable ones from the Internet. On the other hand, we have selected two popular datasets in the human activity recognition field as target domain where the videos are recoded by fixed cameras.

Source Domain Datasets Three public and extensive datasets, HMDB51 [7], OlympicSprots [8] and Virat Release 2.0 [9], are used as source domain. They include a high variability of movements in several locations. The three datasets combined have 79 different activity classes extracted from Youtube, movies or surveillance cameras in 7878 video clips.

Target Domain Datasets The Weizmann dataset [10] is composed by 93 low-resolution (180 x 144, 50 fps) video sequences showing nine different people, each performing 10 natural activities. The KTH dataset [11] has been captured in 4 different scenarios where static cameras have recorded, at low-resolution (160 x 120, 25 fps), 25 subjects performing several times six types of activities.

All videos are processed by means of the state-of-the-art IDT[1] extractor. From the IDTs extracted from the Source Domain datasets, 100000 are randomly selected and used for the GMM training, obtaining 5000 Gaussians, which represent the UBM vocabulary.

B. Results

Considering every sequence alone, a ddCRP is obtained as described in Section III. The number of tables (states) are automatically determined by the ddCRP. This number variates from 4 to 12 depending on the length of the observation sequence. Therefore, we have selected a value of $T = 10$ for the PPK algorithm described in Section IV. The ρ parameter has been chosen $1/2$ so, the PPK becomes the classic Bhattacharyya affinity metric between two probability distributions.

The experiments are conducted using one-subject-out model, where one training sequence per class is randomly selected from the remaining subjects. The result per subject

[1]IDT descriptor code can be downloaded in http://lear.inrialpes.fr/people/wang/download/improved_trajectory_release.tar.gz

TABLE I
ONE-SHOT LEARNING WITH ONE EXAMPLE PER CLASS.

	Weizmann	KTH
ddCPR + PPK (ours)	**79.7%**	65.1%
FO-HMM [12]	68.1%	**67.1%**
Seo and Milanfar [13]	75%	65%
pseudo-marginal likelihood (18)	71.4%	-

are the average of 100 runs and the final result is the average of all subjects.

Table I shows the results for the ddCRP plus PPK in comparison with some results found in the literature. As it can be noticed, our approach seems to perform at the state of the art level, being significantly better for the Weizmann dataset. In the last row of Table I we show the result obtained following the pseudo-marginal likelihood (18) for the Weizmann dataset. We have not been able to get any result for the KTH dataset due to computational restrictions as the computation of this likelihood takes some weeks in a single core machine. We cannot conclude that our approach is better than the pseudo-marginal likelihood, but we can state that from a computational point of view this last approach is not useful when the sequence of observations is large. The computational cost involved in the comparison of two distributions using the PPK is one order of magnitude lower than the one using the pseudo-marginal likelihood.

VI. CONCLUSIONS

This paper tackles with the problem of sequence recognition when the number of samples is scarce, in the limit just one. Therefore, we propose the creation of a UBM of features using videos of public datasets applied to the activity encoding and an unsupervised modelling of the activities with a ddCPR where the temporal information of the incoming observation is taken into account in order to sit customers to tables (i.e., to link observations to hidden states).

It is true that more experimentation has to be done in order to evaluate our method. Nevertheless, preliminary results show it performs at the state of the art for one-shot learning of temporal sequences.

Besides more experimentation, there is another relevant issue to be addressed. In this approach we have mapped temporal features (obtained by a sliding window) into distributions and, once the ddCRP is obtained, a HMM is get from this process in order to use the PPK. However, Dirichlet distributions have been used before to provide an infinite HMM (iHMM) which could be a better strategy for later using the PPK.

There is another interesting topic to try with ddCRP. The above results show that this approach works well with limited data, but their behaviour, when more training examples are available, should be analysed and check how good it is in relation to other approaches.

ACKNOWLEDGMENT

This work was partially supported by Spanish Grant TIN2013- 45312-R (MINECO), Gobierno de Aragon and the European Social Found.

REFERENCES

[1] D. M. Blei and P. I. Frazier, "Distance dependent chinese restaurant processes," *J. Mach. Learn. Res.*, vol. 12, pp. 2461–2488, Nov. 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1953048.2078184

[2] L. Wang and D. Dunson, "Vfast bayesian inference in dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, vol. 20, pp. 196–216, 2011.

[3] T. Jebara, R. Kondor, and A. Howard, "Probability product kernels," *J. Mach. Learn. Res.*, vol. 5, pp. 819–844, Dec. 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=1005332.1016786

[4] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *IEEE Intenational Conference on Computer Vision (ICCV)*, 2013.

[5] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, "Fast unsupervised ego-action learning for first-person sports videos." in *CVPR*. IEEE Computer Society, 2011, pp. 3241–3248.

[6] T. Jebara, Y. Song, and K. Thadani, "Spectral clustering and embedding with hidden markov models." in *ECML*, ser. Lecture Notes in Computer Science, J. N. Kok, J. Koronacki, R. L. de Mntaras, S. Matwin, D. Mladenic, and A. Skowron, Eds., vol. 4701. Springer, 2007, pp. 164–175.

[7] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *IEEE Intenational Conference on Computer Vision (ICCV)*, 2011.

[8] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *European Conference on Computer Vision (ECCV)*, 2010, pp. 392–405.

[9] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai, "A large-scale benchmark dataset for event recognition in surveillance video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, USA, 2011, pp. 3153–3160.

[10] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, December 2007.

[11] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *International Conference on Pattern Recognition*, 2004. [Online]. Available: http://www.nada.kth.se/cvap/actions/

[12] C. Orrite, M. Rodriguez, and M. Montañes, "One-sequence learning of human actions." in *Human Behavior Unterstanding*, A. Salah and B. Lepri, Eds., vol. 7065. Springer Berlin / Heidelberg, 2011, pp. 40–51.

[13] H. J. Seo and P. Milanfar, "Action recognition from one example," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 867–882, 2011. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2010.156