Variational Learning of Beta-Liouville Hidden Markov Models for Infrared Action Recognition

Samr Ali
Electrical and Computer Engineering Department
Concordia University, Montreal, Canada
al_samr@encs.concordia.ca

Nizar Bouguila
Concordia Institute for Information Systems Engineering
Concordia University, Montreal, Canada
nizar.bouguila@encs.concordia.ca

Abstract

Infrared (IR) images are characterized by a lower sensitivity to lighting conditions than the visible spectrum. This opens the door to relatively untapped research potential of automatic recognition systems that are robust to shadows and variability in illumination levels or appearance. IR action recognition (AR) is one such application. It remains a fairly unexplored domain in IR. As such, in this paper, we propose the use of hidden Markov models (HMM) for IR AR. We also derive the mathematical model for the variational learning of Beta-Liouville (BL) HMMs. Next, we present the results of the proposed model on the Infrared Action Recognition (InfAR) dataset. To the best of our knowledge, this is the first application of HMMs to AR in the IR domain, and the first application of the BL HMMs to AR. Experimental results demonstrate promising results using different features extracted from the InfAR dataset.

1. Introduction

Many ubiquitous applications rely on automatic action recognition (AR). These include video surveillance [19], video retrieval [29], and video labeling [19]. As such, AR research has received much attention in recent years. Typically, classification of a given video or image sequence or its assignment to a set of predefined classes is the objective of automatic AR [26]. The task is then based on lower level processing stages such as tracking and segmentation [21].

Different approaches for AR have been studied throughout the years with significant advances made in the past decades [17]. Most of the developed AR approaches are tested and implemented for the visible spectrum [34]. Moreover, an abundant number of visible spectrum AR datasets is available such as KTH [32], Weizmann [15], and UCF101 [33]. Indeed, AR in general is fairly well-studied in the visible light spectrum with multiple successful applications [23].

Nonetheless, many challenges persist that limit the accuracy of AR. One known issue is the intrinsic within-class variability where an individual may carry out the same action differently [26]. Moreover, AR in the visible spectrum still suffers from multiple shortcomings including its high sensitivity to shadow, background clutter, occlusion, and changes in illumination [14].

Thermal infrared (IR) cameras are robust to the aforementioned challenging factors [16]. In particular, humans can be captured in IR under poor illumination conditions, such as in dim light or at night, which is almost impossible to perform in the visible spectrum. Furthermore, robustness to shadow, background clutter, and occlusion challenges are due to their relatively lower temperatures in IR. Hence, IR AR is a promising research field that can potentially outperform applications utilizing the visible light spectrum [23].

A hidden Markov model (HMM) [28] is one of the machine learning approaches that may be used for IR AR. Early works mostly focused on the use of HMMs for discrete and Gaussian data [27]. A primary area of HMM research lies in modeling state emission probabilities of proportional data, i.e. strictly positive data that sum up to one. Multivariate proportional time series data naturally result from numerous preprocessing procedures, such as the commonly used histograms, and occur in various pattern recognition domains.

Applying a Gaussian-based HMM in such a case is not ideal. Indeed, the symmetric property of the Gaussian dis-

tribution and its unbounded support lead to a sub-optimal model of the observations. Nevertheless, this remains the commonplace practice such as in [26]. Enhanced results can then be obtained using asymmetric emission probability distributions with compact support. Moreover, the large applicability of HMMs in a variety of domains motivated the adaption of the learning equations to continuous non-Gaussian data types. These include proportional data [7, 5, 6, 8], and Student's t data [4]. In this paper, we propose the use of the Beta-Liouville (BL) distribution for proportional data modeling using HMMs [10, 12, 2, 9].

Furthermore, a HMM is usually trained with the Baum-Welch method; a variation of the Expectation Maximization algorithm. In this paper, the proposed HMM is trained using a variational learning approach which incorporates prior knowledge into the training process [11]. Employing a variational Bayesian inference technique is advantageous as it overcomes the drawbacks of the Baum Welch algorithm. These include over-fitting or underfitting and sub-optimal generalization performance [11].

The contributions of this paper are threefold: (i) we propose the first mathematical model for the variational learning of BL HMM; (ii) we propose the first BL HMM-based IR AR; (iii) we present the first HMM-based results on the InfAR dataset to the best of our knowledge.

The rest of this paper is organized as follows. Section 2 details the proposed model. Section 3 discusses the experimental results. Finally, Section 4 concludes the paper and briefly presents future work.

2. Variational Learning of the Beta-Liouville Hidden Markov Model

A HMM is characterized by an underlying stochastic process with K hidden states that form a Markov chain. Each of the states is governed by an initial probability π, and the transition between the states at time t can be visualized with a transition matrix $B = \{b_{ii'} = P(s_t = i'|s_{t-1} = i)\}$. In each state s_t, an observation is emitted corresponding to its distribution which may be discrete or continuous. This is the observable stochastic process set.

The emission matrix of the discrete observations can be denoted by $\Xi = \{\Xi_i(m) = P(O_t = \xi_m|s_t = i)\}$ where $[m, t, i] \in [1, M] \times [1, T] \times [1, K]$, and the set of all possible discrete observations $\Xi = \{\xi_1, ..., \xi_m, ..., \xi_M\}$. On the other hand, the respective parameters of a probability distribution define the observation emission for a continuous observed symbol sequence. The Gaussian distribution is the most commonly used and is defined by its mean and covariance matrix $\varkappa = (\mu, \Sigma)$ [27, 30, 38]. Consequently, a mixing matrix must be defined $C = \{c_{ij} = P(m_t = j|s_t = i)\}$ in the case of continuous HMM emission probability distribution where $j \in [1, M]$ such that M is the number of mixture components in set $L = \{m_1, ..., m_M\}$. Hence, a discrete or continuous HMM may be defined with the following respective parameters $\lambda = \{B, \Xi, \pi\}$ or $\{B, C, \varkappa, \pi\}$. In this work, we consider the latter case which is defined as a proportional mixture model of BL distribution.

In D dimensions, a BL distribution is defined as:

$$BL(\vec{x}|\vec{\alpha}, \alpha, \beta) = \frac{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \prod_{d=1}^{D} \frac{x_d^{\alpha_d - 1}}{\Gamma(\alpha_d)} \times \left(\sum_{d=1}^{D} x_d\right)^{\alpha - \sum_{d=1}^{D} \alpha_d} \left(1 - \sum_{d=1}^{D} x_d\right)^{\beta - 1} \quad (1)$$

where $\vec{\alpha} = (\alpha_1, ..., \alpha_D)$, α, and β are the real and strictly positive parameters of the BL distribution, $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$ is the Gamma function, and $\vec{x}$ is a D dimensional vector whereby $\vec{x} \in \mathbb{R}_+^D$ and $\sum_{d=1}^{D} x_d < 1$. For simplification, we also denote $\Lambda = [\vec{\alpha}, \alpha, \beta]$; the parameters of the BL distribution.

Consequently, the likelihood of X, a time-series or sequence of observations of length T, given the model is expressed as:

$$p(X|B, C, \pi, \Lambda) = \sum_{S} \sum_{L} \pi_{s1} \left[\prod_{t=2}^{T} b_{s_{t-1}, s_t}\right] \times \left[\prod_{t=1}^{T} c_{s_t, m_t} p(x_t|\Lambda_{s_t, m_t})\right] \quad (2)$$

where $\Lambda_{ij} = (\Lambda_{1ij}, ..., \Lambda_{Dij})$ with $i \in [1, K]$ where K is the number of states in S; the set of hidden states, and $j \in [1, M]$ where M is the number of mixture components in L; the set of the components of the mixture. M is assumed to be uniform for all the states. Hence, the model is derived for a unique observation sequence for simplification purposes. To consider further observation sequences, an addition of a summation of these sequences would be logically required in the corresponding observation data equations. Furthermore, when $A > 1$, the parameter T is then dependent on each of the available time series observation sequences $\{X^a\}_{a=1,...,A}$ such that it would be denoted as T_a. It is also noteworthy to mention that such a setup is highly recommended since it prevents overfitting.

The exact computation of Equation (2) is intractable due to the need of summation over all possible combinations of mixture components and states. Consequently, the typical methodology for its solution constitutes of the maximization of the data likelihood with respect to the parameters of the model using the Baum-Welch algorithm [27]. Nonetheless, this approach suffers from several drawbacks. These include overfitting and absence of a convergence guarantee due to the general multimodal nature of the data likelihood function.

On the other hand, an estimation of the model may be derived using the variational Bayesian approach. This uses the posterior probabilities through the assignment of parameter priors for integrating out the marginal likelihood of the data. Hence, all the model parameters are regarded as random variables. The complete data likelihood is then denoted as:

$$p(X) = \int d\pi dB dC d\Lambda \sum_{S,L} p(B, C, \pi, \Lambda) \; p(X, S, L|B, C, \pi, \Lambda) \quad (3)$$

Equation (3) is still computationally intractable. This is due to the exponential growth of the number of possible sequences to be summed as the length of the time series increases [20]. However, an introduction of the approximate distribution $q(B, C, \pi, \Lambda, S, L)$ of the true posterior $p(B, C, \pi, \Lambda, S, L|X)$ enables us to derive a lower bound. Thus, using Jensen's inequality and Equation (3), the lower bound can be expressed as:

$$\begin{aligned}\ln(p(X)) &= \ln\left\{\int d\pi dB dC d\Lambda \sum_{S,L} p(B, C, \pi, \Lambda) \times p(X, S, L|B, C, \pi, \Lambda)\right\} \\ &\geq \int d\pi dB dC d\Lambda \sum_{S,L} q(B, C, \pi, \Lambda, S, L) \\ &\times \ln\left\{\frac{p(B, C, \pi, \Lambda) p(X, S, L|B, C, \pi, \Lambda)}{q(B, C, \pi, \Lambda, S, L)}\right\} \quad (4)\end{aligned}$$

When q is equal the true posterior, the inequality is tight. Hence,

$$\begin{aligned}&\ln(p(X)) = \mathcal{L}(q) \\ &- KL(q(B, C, \pi, \Lambda, S, L)||p(B, C, \pi, \Lambda, S, L|X)) \quad (5)\end{aligned}$$

where $\mathcal{L}(q)$ is the lower bound and KL is the Kullback-Leibler distance between the true posterior and the approximate distribution [11, 8].

The computation of the exact posterior distribution is intractable, so we only account for a certain family of distributions. As per the studied assumptions in [11, 8, 3, 20, 25], q may be factorized, i.e. $q(B, C, \pi, \Lambda, S, L) = q(B)q(C)q(\pi)q(\Lambda)q(S, L)$ where $q(\Lambda) = q(\vec{\alpha})q(\alpha)q(\beta)$, with a similar factorization applying to p. $\mathcal{L}(q)$ can then be expressed as:

$$\begin{aligned}&\ln(p(X)) \geq \sum_{S,L} \int dB dC d\pi d\vec{\alpha} d\alpha d\beta q(B)q(C)q(\pi) \times \\ &q(\vec{\alpha})q(\alpha)q(\beta)q(S, L)\{\ln(p(\pi)) + \ln(p(B)) + \ln(p(C)) + \\ &\ln(p(\vec{\alpha})) + \ln(p(\alpha)) + \ln(p(\beta)) + \ln(p(\pi_{s_1})) + \sum_{t=2}^{T} \ln(b_{s_{t-1}, s_t}) \\ &+ \sum_{t=1}^{T} \ln(c_{s_t, m_t}) + \sum_{t=1}^{T} \ln(p(x_t|\vec{\alpha}_{s_t, m_t}, \alpha_{s_t, m_t}, \beta_{s_t, m_t})) - \\ &\ln(q(S, L)) - \ln(q(\pi)) - \ln(q(B)) - \ln(q(C)) - \ln(q(\vec{\alpha})) - \\ &\ln(q(\alpha)) - \ln(q(\beta)) = F(q(\pi)) + F(q(B)) + F(q(C)) \\ &+ F(q(\vec{\alpha})) + F(q(\alpha)) + F(q(\beta)) + F(q(S, L))\} \quad (6)\end{aligned}$$

In general, there are multiple maxima to the above lower bound; i.e. it is not convex. This implies that the solution is dependent on the initialization. The priors of the parameters must then be defined to evaluate Equation (6). Since the coefficients of the parameters π, B, and C are all less than one, strictly positive, and with a sum result equal to one for each row summation, their priors are chosen as Dirichlet distributions as follows:

$$\begin{aligned}p(\pi) &= \mathcal{D}(\pi|\phi^{\pi}) = \mathcal{D}(\pi_1, ..., \pi_K|\phi_1^{\pi}, ..., \phi_K^{\pi}), \\ p(B) &= \prod_{i=1}^{K} \mathcal{D}(b_{i_1}, ..., b_{i_K}|\phi_{i_1}^{B}, ..., \phi_{i_K}^{B}), \\ p(C) &= \prod_{i=1}^{M} \mathcal{D}(c_{i_1}, ..., c_{i_M}|\phi_{i_1}^{C}, ..., \phi_{i_M}^{C}) \quad (7)\end{aligned}$$

Similarly, a conjugate prior must also be defined over the BL parameters $\vec{\alpha}$, α, and β. We adopt the Gamma distribution $\mathcal{G}(.)$ for positive conjugate prior approximations of the latter parameters as previously investigated by in [11]. Hence, we define the prior distributions as:

$$p(\{\vec{\alpha}\}_{i,j,l=1}^{K,M,D}) = \prod_{i=1}^{K} \prod_{j=1}^{M} \prod_{l=1}^{D} \mathcal{G}(\alpha_{ijl}|u_{ijl}, v_{ijl}), \quad (8)$$

$$p(\{\alpha\}_{i,j=1}^{K,M}) = \prod_{i=1}^{K} \prod_{j=1}^{M} \mathcal{G}(\alpha_{ij}|g_{ij}, h_{ij}), \quad (9)$$

$$p(\{\beta\}_{i,j=1}^{K,M}) = \prod_{i=1}^{K} \prod_{j=1}^{M} \mathcal{G}(\beta_{ij}|e_{ij}, r_{ij}) \quad (10)$$

where the hyperparameters u, g, h, e, r, and v are strictly positive.

The iterative variational Bayesian inference process consists of two alternating steps; the E-step and the M-step. All of the parameters of the model are then learned through a

sequential repetition of a M-step followed by an E-step until convergence. Hidden states and mixture components are updated in the M-step, so all (S, L) terms in Equation (6) are not considered. On the other hand, $q(S, L)$ is subsequently updated in the E-step; now keeping all other parameters fixed.

The following optimizations of $q(B)$, $q(C)$, and $q(\pi)$ are applicable to other continuous HMMs as they are independent of the emission distribution used. Therefore, these have already been studied in [3, 8]. As such, only the main equations are given and the reader is referred to the aforementioned references for further details. Consequently, the derivation of the equations with terms pertaining only to the B parameter from Equation (6) gives:

$$F(q(B)) = \int q(B) \ln \left[\frac{\prod_{i=1}^{K} \prod_{j=1}^{K} b_{ij}^{\omega_{ij}^{B}-1}}{q(B)} \right] dB \quad (11)$$

with

$$\omega_{ij}^{B} = \sum_{t=2}^{T} \gamma_{ijt}^{B} + \phi_{ij}^{B} \quad (12)$$

and

$$\gamma_{ijt}^{B} \triangleq q(s_{t-1} = i, s_t = j) \quad (13)$$

where γ_{ijt}^{B} is a local probability typically computed with a forward-backward algorithm in a HMM framework [27]. To maximize $F(q(B))$, we apply the Gibbs inequality which results in:

$$q(B) = \prod_{i=1}^{K} \mathcal{D}(a_{i1}, ..., a_{iK} | \omega_{i1}^{B}, ..., \omega_{iK}^{B}) \quad (14)$$

Similarly for the π parameter:

$$q(\pi) = \mathcal{D}(\pi_1, ..., \pi_K | \omega_1^{\pi}, ..., \omega_K^{\pi}) \quad (15)$$

with

$$\omega_i^{\pi} = \gamma_i^{\pi} + \phi_i^{\pi} \quad (16)$$

and

$$\gamma_i^{\pi} \triangleq q(s_1 = i) \quad (17)$$

Finally, for the C parameter:

$$q(C) = \prod_{i=1}^{K} \mathcal{D}(c_{i1}, ..., c_{iM} | \omega_{i1}^{C}, ..., \omega_{iM}^{C}) \quad (18)$$

with

$$\omega_{ij}^{C} = \sum_{t=1}^{T} \gamma_{ijt}^{C} + \phi_{ij}^{C} \quad (19)$$

and

$$\gamma_{ijt}^{C} \triangleq q(s_t = i, m_t = j) \quad (20)$$

Next, we tackle the optimization of $F(q(\Lambda))$. From Equation (6), we obtain:

$$F(q(\Lambda)) = \int q(\Lambda) \ln \left\{ \frac{\prod_{i=1}^{K} \prod_{j=1}^{M} p(\Lambda_{ij}) \prod_{t=1}^{T} p(x_t \Lambda_{ij})^{\gamma_{ijt}^{C}}}{q(\Lambda)} \right\} d\Lambda \quad (21)$$

In order to achieve tractability, we apply the previously discussed factorial approximation of $q(\Lambda)$ as in [1]. We note that the solution thus far is presented corresponding to that of a finite BL mixture model as investigated in [10]. This leads to the following evaluations:

$$q(\vec{\alpha}) = \prod_{l=1}^{D} \prod_{i=1}^{K} \prod_{j=1}^{M} \mathcal{G}(\alpha_{ijl} | u_{ijl}^{*}, v_{ijl}^{*}) \quad (22)$$

$$q(\alpha) = \prod_{i=1}^{K} \prod_{j=1}^{M} \mathcal{G}(\alpha_{ij} | g_{ij}^{*}, h_{ij}^{*}) \quad (23)$$

$$q(\beta) = \prod_{i=1}^{K} \prod_{j=1}^{M} \mathcal{G}(\beta_{ij} | e_{ij}^{*}, r_{ij}^{*}) \quad (24)$$

where

$$u_{ijl}^{*} = u_{ijl} + \sum_{p=1}^{P} \langle Z_{pij} \rangle \bar{\alpha}_{ijl} \left[\Psi \left(\sum_{d=1}^{D} \bar{\alpha}_{ijd} \right) - \Psi(\bar{\alpha}_{ijl}) + \sum_{d=1, d \neq l}^{D} \Psi' \left(\sum_{d=1}^{D} \bar{\alpha}_{ijd} \right) \bar{\alpha}_{ijd} (\langle \ln(\alpha_{ijd}) \rangle - \ln(\bar{\alpha}_{ijd})) \right] \quad (25)$$

$$v_{ijl}^{*} = v_{ijl} - \sum_{p=1}^{P} \langle Z_{pij} \rangle \left[\ln(X_{pl}) - \ln \left(\sum_{d=1}^{D} X_{pd} \right) \right] \quad (26)$$

$$g_{ij}^{*} = g_{ij} + \sum_{p=1}^{P} \langle Z_{pij} \rangle [\Psi(\bar{\alpha}_{ij} + \bar{\beta}_{ij}) - \Psi(\bar{\alpha}_{ij}) + \bar{\beta}_{ij} \Psi'(\bar{\alpha}_{ij} + \bar{\beta}_{ij})(\langle \ln(\beta_{ij}) \rangle - \ln(\bar{\beta}_{ij}))] \bar{\alpha}_{ij} \quad (27)$$

$$h_{ij}^{*} = h_{ij} - \sum_{p=1}^{P} \langle Z_{pij} \rangle \ln \left(\sum_{d=1}^{D} X_{pd} \right) \quad (28)$$

$$e_{ij}^{*} = e_{ij} + \sum_{p=1}^{P} \langle Z_{pij} \rangle [\Psi(\bar{\alpha}_{ij} + \bar{\beta}_{ij}) - \Psi(\bar{\beta}_{ij}) + \bar{\alpha}_{ij} \Psi'(\bar{\alpha}_{ij} + \bar{\beta}_{ij})(\langle \ln(\alpha_{ij}) \rangle - \ln(\bar{\alpha}_{ij}))] \bar{\beta}_{ij} \quad (29)$$

$$r_{ij}^{*} = r_{ij} - \sum_{p=1}^{P} \langle Z_{pij} \rangle \ln \left(1 - \sum_{d=1}^{D} X_{pd} \right) \tag{30}$$

with i and j fixed for P observation vectors where $l \in [1, D]$, $i \in [1, K]$, and $j \in [1, M]$. $\Psi(.)$ is the digamma function, and $\Psi'(.)$ is the trigamma function; the logarithmic first and second derivatives of the Gamma function respectively. The $*$ superscript implies the optimization of each of the corresponding parameters that the symbol is presented upon and $\langle . \rangle$ denotes the expectation with respect to the optimized parameter, accordingly. Moreover, $Z_{pij} = 1$ if X_{pt} belongs to state i and mixture component j and $Z_{pij} = 0$ otherwise, i.e. it is an indicator function. Then, the weights of the data samples with respect to each mixture component are defined within the HMM framework. These are also known as the responsibilities. Consequently, $\langle Z_{pij} \rangle = \sum_{t=1}^{T} \gamma_{pijt}^{C} = p(s = i, m = j|X)$ and the responsibilities are computed via the forward-backward algorithm [27]. The definitions of the expected values of the parameters in the aforementioned equations are as follows:

$$\bar{\alpha}_{ijl} = \frac{u_{ijl}^{*}}{v_{ijl}^{*}}, \bar{\alpha}_{ij} = \frac{g_{ij}^{*}}{h_{ij}^{*}}, \bar{\beta}_{ij} = \frac{e_{ij}^{*}}{r_{ij}^{*}} \tag{31}$$

$$\langle \ln(\alpha_{ijl}) \rangle = \Psi(u_{ijl}^{*}) - \ln(v_{ijl}^{*}) \tag{32}$$

$$\langle \ln(\alpha_{ij}) \rangle = \Psi(g_{ij}^{*}) - \ln(h_{ij}^{*}) \tag{33}$$

$$\langle \ln(\beta_{ij}) \rangle = \Psi(e_{ij}^{*}) - \ln(r_{ij}^{*}) \tag{34}$$

This concludes the M-step of the algorithm. $q(S, L)$ is then estimated in the E-step with the previously evaluated parameters now fixed. Equation (6) can be rearranged as studied in [8] to:

$$\mathcal{L}(q) = F(q(S, L)) - KL(q(B, C, \pi, \Lambda)||p(B, C, \pi, \Lambda)) \tag{35}$$

where

$$\begin{aligned} F(q(S, L)) = & \sum_{S} q(S) \int q(\pi) \ln(\pi_{s1}) d\pi + \\ & \sum_{S} q(S) \int q(B) \sum_{t=2}^{T} \ln(b_{s_{t-1}, s_t}) dB + \\ & \sum_{S,L} q(S, L) \int q(C) \sum_{t=1}^{T} \ln(c_{s_t, m_t}) dC + \\ & \sum_{S,L} q(S, L) \int q(\Lambda) \sum_{t=1}^{T} \ln(p(x_t|\vec{\alpha}_{s_t, m_t}, \alpha_{s_t, m_t}, \beta_{s_t, m_t})) d\Lambda + \\ & - \sum_{S,L} q(S, L) \ln(q(S, L)), \end{aligned} \tag{36}$$

and we naturally define:

$$\begin{aligned} \pi_i^{*} &\triangleq \exp\left[\langle \ln(\pi_i) \rangle_{q(\pi)}\right], \\ \pi_i^{*} &= \exp\left[\Psi(\omega_i^{\pi}) - \Psi(\sum_i \omega_i^{\pi})\right], \\ b_{jj'}^{*} &\triangleq \exp\left[\langle \ln(b_{jj'}) \rangle_{q(B)}\right], \\ b_{jj'}^{*} &= \exp\left[\Psi(\omega_{jj'}^{B}) - \Psi(\sum_{j'} \omega_{jj'}^{B})\right], \\ c_{ij}^{*} &\triangleq \exp\left[\langle \ln(c_{ij}) \rangle_{q(C)}\right], \\ c_{ij}^{*} &= \exp\left[\Psi(\omega_{ij}^{C}) - \Psi(\sum_{j} \omega_{ij}^{C})\right] \end{aligned} \tag{37}$$

The final optimization that needs to be performed is:

$$\begin{aligned} & \ln(p^{*}(X_t|\vec{\alpha}_{s_t, m_t}, \alpha_{s_t, m_t}, \beta_{s_t, m_t})) = \\ & \int q(\Lambda) \ln(p(X_t|\vec{\alpha}_{s_t, m_t}, \alpha_{s_t, m_t}, \beta_{s_t, m_t})) d\Lambda, \end{aligned} \tag{38}$$

where

$$\begin{aligned} & p(X_t|\vec{\alpha}_{s_t, m_t}, \alpha_{s_t, m_t}, \beta_{s_t, m_t}) = \\ & \left[\frac{\Gamma(\sum_{d=1}^{D} \alpha_{ijd}) \Gamma(\alpha_{ij} + \beta_{ij})}{\Gamma(\alpha_{ij}) \Gamma(\beta_{ij})} \prod_{d=1}^{D} \frac{X_{td}^{\alpha_{ijd} - 1}}{\Gamma(\alpha_{ijd})} \times \right. \\ & \left. \left(\sum_{d=1}^{D} X_{td} \right)^{\alpha_{ij} - \sum_{d=1}^{D} \alpha_{ijd}} \left(1 - \sum_{d=1}^{D} X_{td} \right)^{\beta_{ij} - 1} \right]^{\gamma_{ijt}^{C}} \end{aligned} \tag{39}$$

We then substitute Equation (39) in Equation (38) and breakdown the distribution $BL(\vec{x}|\vec{\alpha}, \alpha, \beta)$ to a product decomposition corresponding to the prior factorization assumption made to $q(\Lambda)$. This yields the following evaluation:

$$\begin{aligned} & \ln(p^{*}(X_t|\vec{\alpha}_{s_t, m_t}, \alpha_{s_t, m_t}, \beta_{s_t, m_t})) = \gamma_{ijt}^{C} \int q(\vec{\alpha}) q(\alpha, \beta) \\ & \ln(\nu(X_t|\vec{\alpha}_{s_t, m_t}) \eta(X_t|\alpha_{s_t, m_t}, \beta_{s_t, m_t})) d\vec{\alpha} d\alpha d\beta \\ & = \gamma_{ijt}^{C} \left(\langle \ln(\nu(X_t|\vec{\alpha})) \rangle_{q(\vec{\alpha})} + \langle \ln(\eta(X_t|\alpha, \beta)) \rangle_{q(\alpha, \beta)} \right) \end{aligned} \tag{40}$$

where

$$\langle \ln(\nu(X_t|\vec{\alpha}))\rangle_{q(\vec{\alpha})} = \left\langle \ln\left(\frac{\Gamma(\sum_{d=1}^{D}\alpha_{ijd})}{\prod_{d=1}^{D}\Gamma(\alpha_{ijd})}\right)\right\rangle_{q(\vec{\alpha})} + \sum_{d=1}^{D}\ln(X_{td})\langle\alpha_{ijd}-1\rangle_{q(\vec{\alpha})} - \ln\left(\sum_{d=1}^{D}X_{td}\right)\sum_{d=1}^{D}\langle\alpha_{ijd}\rangle_{q(\vec{\alpha})}$$
$$= J(\alpha_{ijl}) + \sum_{d=1}^{D}\ln(X_{td})\left(\frac{u_{ijd}}{v_{ijd}}-1\right) - \ln\left(\sum_{d=1}^{D}X_{td}\right)\times \sum_{d=1}^{D}\left(\frac{u_{ijd}}{v_{ijd}}\right) \quad (41)$$

and

$$\langle \ln(\eta(X_t|\alpha_{ij},\beta_{ij}))\rangle_{q(\alpha_{ij},\beta_{ij})} = \left\langle \ln\left(\frac{\Gamma(\alpha_{ij}+\beta_{ij})}{\Gamma(\alpha_{ij})\Gamma(\beta_{ij})}\right)\right\rangle_{q(\alpha,\beta)} + \ln\left(\sum_{d=1}^{D}X_{td}\right)\langle\alpha_{ij}\rangle_{q(\alpha,\beta)} + \ln\left(1-\sum_{d=1}^{D}X_{td}\right)\langle\beta_{ij}-1\rangle_{q(\alpha,\beta)}$$
$$= J(\alpha_{ij},\beta_{ij}) + \ln\left(\sum_{d=1}^{D}X_{td}\right)\left(\frac{g_{ij}}{h_{ij}}\right) + \ln\left(1-\sum_{d=1}^{D}X_{td}\right)\left(\frac{e_{ij}}{r_{ij}}-1\right) \quad (42)$$

$J(\alpha_{ijl})$ and $J(\alpha_{ij},\beta_{ij})$ are analytically intractable. Consequently, they are approximated by their lower bounds as derived in [11]. Using the second order Taylor approximation method, $J(\alpha_{ijl})$ and $J(\alpha_{ij},\beta_{ij})$ are then denoted as follows:

$$J(\alpha_{ijl}) \geq \ln\left(\frac{\Gamma(\sum_{d=1}^{D}\bar{\alpha}_{ijd})}{\prod_{d=1}^{D}\Gamma(\bar{\alpha}_{ijd})}\right) + \sum_{d=1}^{D}\bar{\alpha}_{ijd}\times \left[\Psi\left(\sum_{l=1}^{D}\bar{\alpha}_{ijl}\right) - \Psi(\bar{\alpha}_{ijd})\right][\langle\ln(\alpha_{ijd})\rangle - \ln(\bar{\alpha}_{ijd})] + \frac{1}{2}\sum_{d=1}^{D}\bar{\alpha}_{ijd}^2\left[\Psi'\left(\sum_{l=1}^{D}\bar{\alpha}_{ijl}\right) - \Psi'(\bar{\alpha}_{ijd})\right]\times \langle(\ln(\alpha_{ijd})-\ln(\bar{\alpha}_{ijd}))^2\rangle + \frac{1}{2}\sum_{d=1}^{D}\sum_{l=1,l\neq d}^{D}\bar{\alpha}_{ijd}\bar{\alpha}_{ijl}\times \Psi'\left(\sum_{y=1}^{D}\bar{\alpha}_{ijy}\right)(\langle\ln(\alpha_{ijd})\rangle-\ln(\bar{\alpha}_{ijd}))(\langle\ln(\alpha_{ijl})\rangle-\ln(\bar{\alpha}_{ijl})) \quad (43)$$

$$J(\alpha_{ij},\beta_{ij}) \geq \ln\left(\frac{\Gamma(\bar{\alpha}_{ij}+\bar{\beta}_{ij})}{\Gamma(\bar{\alpha}_{ij})\Gamma(\bar{\beta}_{ij})}\right) + \bar{\alpha}_{ij}(\Psi(\bar{\alpha}_{ij}+\bar{\beta}_{ij})-\Psi(\bar{\alpha}_{ij}))(\langle\ln(\alpha_{ij})\rangle-\ln(\bar{\alpha}_{ij})) + \bar{\beta}_{ij}(\Psi(\bar{\alpha}_{ij}+\bar{\beta}_{ij})-\Psi(\bar{\beta}_{ij}))(\langle\ln(\beta_{ij})\rangle-\ln(\bar{\beta}_{ij})) + \frac{1}{2}\bar{\alpha}_{ij}^2(\Psi'(\bar{\alpha}_{ij}+\bar{\beta}_{ij})-\Psi'(\bar{\alpha}_{ij}))\langle(\ln(\alpha_{ij})-\ln(\bar{\alpha}_{ij}))^2\rangle + \frac{1}{2}\bar{\beta}_{ij}^2(\Psi'(\bar{\alpha}_{ij}+\bar{\beta}_{ij})-\Psi'(\bar{\beta}_{ij}))\langle(\ln(\beta_{ij})-\ln(\bar{\beta}_{ij}))^2\rangle + \bar{\alpha}_{ij}\bar{\beta}_{ij}\Psi'(\bar{\alpha}_{ij}+\bar{\beta}_{ij})(\langle\ln(\alpha_{ij})\rangle-\ln(\bar{\alpha}_{ij}))(\langle\ln(\beta_{ij})\rangle-\ln(\bar{\beta}_{ij})) \quad (44)$$

where $\langle(\ln(\alpha_{ijd})-\ln(\bar{\alpha}_{ijd}))^2\rangle = (\Psi(u_{ijd})-\ln(u_{ijd}))^2 + \Psi'(u_{ijd})$, $\langle(\ln(\alpha_{ij})-\ln(\bar{\alpha}_{ij}))^2\rangle = (\Psi(g_{ij})-\ln(g_{ij}))^2 + \Psi'(g_{ij})$, and $\langle(\ln(\beta_{ij})-\ln(\bar{\beta}_{ij}))^2\rangle = (\Psi(e_{ij})-\ln(e_{ij}))^2 + \Psi'(e_{ij})$ as derived in [24].

Finally, by substituting Equation (43) into Equation (41), Equation (44) into Equation (42), and Equation (37) into Equation (36), we yield:

$$F(q(S,L)) = \sum_{S,L}q(S,L)\ln\left(p^*(X_t|\vec{\alpha}_{s_t,m_t},\alpha_{s_t,m_t},\beta_{s_t,m_t})\frac{\pi^*_{s_1}\prod_{t=2}^{T}b^*_{s_{t-1},s_t}\prod_{t=1}^{T}c^*_{s_t,m_t}}{q(S,L)}\right) \quad (45)$$

whereby the optimized $q(S,L)$ can then be denoted as:

$$q(S,L) = \frac{1}{W}\pi^*_{s_1}\prod_{t=2}^{T}b^*_{s_{t-1},s_t}\prod_{t=1}^{T}c^*_{s_t,m_t}\times p^*(X_t|\vec{\alpha}_{s_t,m_t},\alpha_{s_t,m_t},\beta_{s_t,m_t}) \quad (46)$$

where W is a normalizing constant and represents the likelihood of the optimized HMM which can be computed with a forward-backward algorithm [27]. This is defined as:

$$W = \sum_{S,L}\pi^*_{s_1}\prod_{t=2}^{T}b^*_{s_{t-1},s_t}\prod_{t=1}^{T}c^*_{s_t,m_t}p^*(X_t|\vec{\alpha}_{s_t,m_t},\alpha_{s_t,m_t},\beta_{s_t,m_t}) \quad (47)$$

3. Experimental Results

In this section, we present our experimental results of the proposed model on the challenging AR IR dataset, *InfAR* [13]. The dataset consists of 12 action classes with a total of 600 video clips. The average length of the videos is 4 seconds with a frame rate of 25 and a resolution of 293 × 256. Classes of a single person action with 10 video samples each were chosen for the training and testing of the proposed model. This results in a total of seven classes with extracted example images of each of the classes shown in Fig. 1.

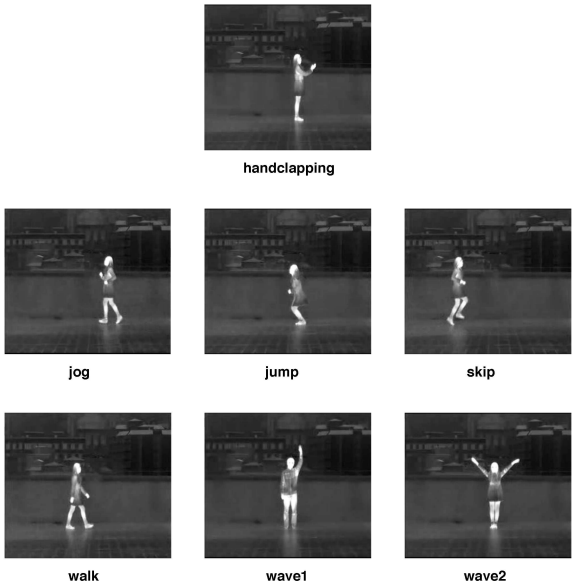

Figure 1. Example sample images from the InfAR dataset.

We represent each of the videos with a series of extracted histogram of optical flow (HOF) and motion boundary histogram (MBH) descriptors which may be detected using any interest point detector [31]. In our experiments, we extract the points along the motion trajectory as in [35]. This set of extracted features represents the training and testing data with a leave-one-out cross validation scheme.

A HMM is then trained for each class using the aforementioned data. For the testing stage, the likelihood of each testing video sequence is calculated by the respective seven trained HMMs and the class label is assigned according to the maximum resulting likelihood. We train a BL HMM with each set of training features for each of the classes 9 times in order to ensure robustness of the methodology. This results in a total of 630 trained HMMs. We report our results as an average across the training times. Our experimental setup can be observed in Fig. 2. It is noteworthy to mention that the number of states and the respective number of mixture components of the proposed BL HMM for this application have been set experimentally to $K = 2$ and $M = 2$ respectively. A similar setup is carried out for a Gaussian-based HMM for comparison of the final classification results.

We achieve superior results of 77.94% when training with the HOF features compared with 42.86% with the Gaussian HMM. Moreover, the average accuracy of proposed model is 89.05% and 92.06% with the horizontal and vertical MBH features respectively versus 85.7% using the benchmark. As such, the proposed HMM clearly outperforms the benchmark and shows promising results. The confusion matrices of the different features with the proposed BL HMM can be observed in Fig. 3, Fig. 4, and Fig.

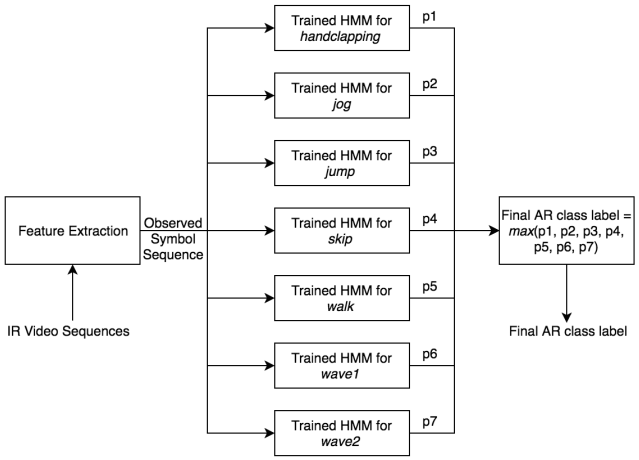

Figure 2. Experimental setup for testing of the proposed trained hidden Markov models (HMM) for infrared (IR) action recognition (AR) classification. p1, p2, p3, p4, p5, p6, p7 are the respective likelihoods of each of the trained HMMs.

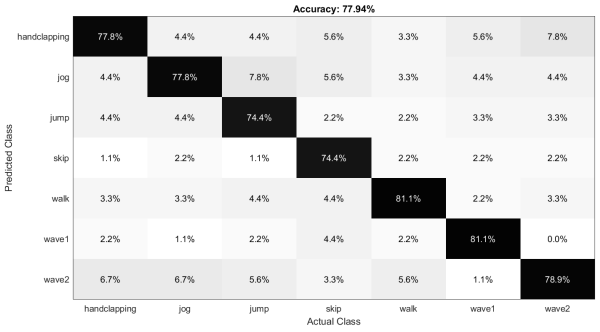

Figure 3. Confusion matrix for BL HMM trained with HOF features.

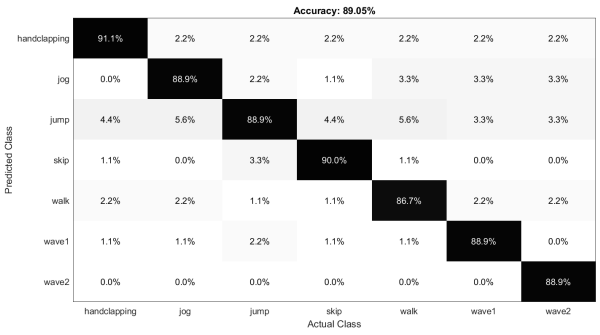

Figure 4. Confusion matrix for BL HMM trained with horizontal MBH features.

5.

Furthermore, our results are comparable to other methods reported in the literature. This includes the the two-stream 3D convolutional neural network (CNN) that achieves 77.50% average precision (AP) [18], the optical flow field 3D CNN with 75.42% AP [18], and 79.25% for the three-stream trajectory-pooled deep-convolutional descriptors methodology in [22]. This is also the case for

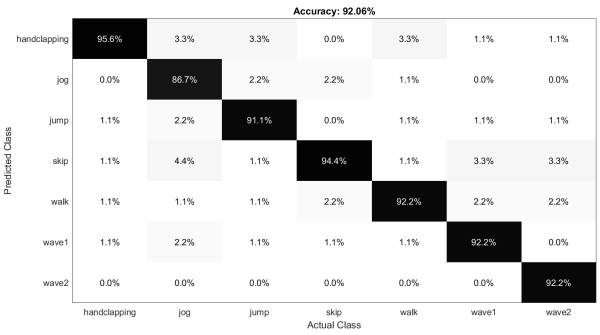

Figure 5. Confusion matrix for BL HMM trained with vertical MBH features.

Table 1. Comparison of the Average Precision (AP) of the proposed BL HMM with other methods in the literature. Results of the proposed model are highlighted in italics.

Method	AP
Two stream 3D CNN [18]	75.42%
Optical flow field 3D CNN [18]	77.50%
Deep-convolutional descriptors [22]	79.25%
HOF [13]	68.58%
Dense trajectories [36]	68.66%
Improved dense trajectories [37]	71.83%
BL HMM (HOF)	*78.41%*
BL HMM (Horizontal MBH)	*89.57%*
BL HMM (Vertical MBH)	*92.29%*

various handcrafted features extracted for the InfAR dataset such as 68.58% for the HOF [13], 68.66% for the dense trajectories [36], and 71.83% for the improved dense trajectories [37]. A comparison of the achieved results with the AP of the proposed model can be observed in Table 1.

4. Conclusion

Robust AR is an active area of research. IR imaging offers a relatively unexplored alternative to the traditional employment of the visible light spectrum for automatic AR systems. It has the advantage of insensitivity to variability in lighting conditions, appearance, and shadows. In this paper, we propose a novel BL-based HMM trained with variational learning and introduce it for IR AR. To the best of our knowledge, this is the first application of HMMs to AR in the IR domain, and the first use of BL HMMs for AR. We achieve state-of-the-art results on the InfAR dataset with the proposed model. Finally, our future plans include investigation of multiperson IR AR and other computer vision applications with the proposed HMM as well as integration of feature selection within our model.

5. Acknowledgments

The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[2] N. Bouguila. Hybrid generative/discriminative approaches for proportional data modeling and classification. *IEEE Trans. on Knowl. and Data Eng.*, 24(12):2184–2202, Dec. 2012.

[3] S. P. Chatzis and D. I. Kosmopoulos. A variational bayesian methodology for hidden markov models utilizing student's-t mixtures. *Pattern Recognition*, 44(2):295–306, 2011.

[4] S. P. Chatzis, D. I. Kosmopoulos, and T. A. Varvarigou. Robust sequential data modeling using an outlier tolerant hidden markov model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1657–1669, Sept 2009.

[5] L. Chen, D. Barber, and J.-M. Odobez. Dynamical dirichlet mixture model. Idiap-RR Idiap-RR-02-2007, IDIAP, 2007.

[6] E. Epaillard and N. Bouguila. Hidden markov models based on generalized dirichlet mixtures for proportional data modeling. In N. El Gayar, F. Schwenker, and C. Suen, editors, *Artificial Neural Networks in Pattern Recognition*, pages 71–82, Cham, 2014. Springer International Publishing.

[7] E. Epaillard and N. Bouguila. Proportional data modeling with hidden markov models based on generalized dirichlet and beta-liouville mixtures applied to anomaly detection in public areas. *Pattern Recognition*, 55:125 – 136, 2016.

[8] E. Epaillard and N. Bouguila. Variational bayesian learning of generalized dirichlet-based hidden markov models applied to unusual events detection. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2018.

[9] W. Fan, F. R. Al-Osaimi, N. Bouguila, and J. Du. Accelerated variational inference for beta-liouville mixture learning with application to 3d shapes recognition. In *International Conference on Control, Decision and Information Technologies, CoDIT 2016, Saint Julian's, Malta, April 6-8, 2016*, pages 394–398, 2016.

[10] W. Fan and N. Bouguila. Learning finite beta-liouville mixture models via variational bayes for proportional data clustering. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 1323–1329. AAAI Press, 2013.

[11] W. Fan and N. Bouguila. Online learning of a dirichlet process mixture of beta-liouville distributions via variational inference. *IEEE Transactions on Neural Networks and Learning Systems*, 24(11):1850–1862, Nov 2013.

[12] W. Fan, N. Bouguila, S. Bourouis, and Y. Laalaoui. Entropy-based variational bayes learning framework for data clustering. *IET Image Processing*, 12(10):1762–1772, 2018.

[13] C. Gao, Y. Du, J. Liu, J. Lv, L. Yang, D. Meng, and A. G. Hauptmann. Infar dataset: Infrared action recognition at dif-

ferent times. *Neurocomputing*, 212:36 – 47, 2016. Chinese Conference on Computer Vision 2015 (CCCV 2015).

[14] C. Gao, Y. Du, J. Liu, L. Yang, and D. Meng. A new dataset and evaluation for infrared action recognition. In *CCF Chinese Conference on Computer Vision*, pages 302–312. Springer, 2015.

[15] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, December 2007.

[16] J. Han and B. Bhanu. Fusion of color and infrared video for moving human detection. *Pattern Recognition*, 40(6):1771–1784, 2007.

[17] S. Herath, M. Harandi, and F. Porikli. Going deeper into action recognition: A survey. *Image and Vision Computing*, 60:4 – 21, 2017. Regularization Techniques for High-Dimensional Data Analysis.

[18] Z. Jiang, V. Rozgic, and S. Adali. Learning spatiotemporal features for infrared action recognition with 3d convolutional neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 309–317. IEEE, 2017.

[19] C.-B. Jin, S. Li, T. D. Do, and H. Kim. Real-time human action recognition using cnn over temporal images for static video surveillance cameras. In Y.-S. Ho, J. Sang, Y. M. Ro, J. Kim, and F. Wu, editors, *Advances in Multimedia Information Processing – PCM 2015*, pages 330–339, Cham, 2015. Springer International Publishing.

[20] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, Nov. 1999.

[21] M. H. Kabir, M. R. Hoque, K. Thapa, and S.-H. Yang. Two-layer hidden markov model for human activity recognition in home environments. *International Journal of Distributed Sensor Networks*, 12(1):4560365, 2016.

[22] Y. Liu, Z. Lu, J. Li, T. Yang, and C. Yao. Global temporal representation based cnns for infrared action recognition. *IEEE Signal Processing Letters*, 25(6):848–852, June 2018.

[23] Y. Liu, Z. Lu, J. Li, C. Yao, and Y. Deng. Transferable feature representation for visible-to-infrared cross-dataset human action recognition. *Complexity*, 2018, 2018.

[24] Z. Ma and A. Leijon. Bayesian estimation of Beta mixture models with variational inference. *IEEE Trans. Pattern Anal. Mach. Intel.*, 33(11):2160–2173, 2011.

[25] D. J. C. MacKay. Ensemble Learning for Hidden Markov Models. *Technical Report*, (1995):0–6, 1997.

[26] Z. Moghaddam and M. Piccardi. Training initialization of hidden markov models in human action recognition. *IEEE Transactions on Automation Science and Engineering*, 11(2):394–408, April 2014.

[27] L. Rabiner and B. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4–16, Jan 1986.

[28] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[29] M. Ramezani and F. Yaghmaee. A review on human action analysis in videos for retrieval applications. *Artif. Intell. Rev.*, 46(4):485–514, Dec. 2016.

[30] M. Rodriguez, C. Orrite, C. Medrano, and D. Makris. One-shot learning of human activity with an map adapted gmm and simplex-hmm. *IEEE Transactions on Cybernetics*, 47(7):1769–1780, July 2017.

[31] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *Int. J. Comput. Vision*, 37(2):151–172, June 2000.

[32] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03*, ICPR '04, pages 32–36, Washington, DC, USA, 2004. IEEE Computer Society.

[33] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[34] M. Vrigkas, C. Nikou, and I. A. Kakadiaris. A review of human activity recognition methods. *Frontiers in Robotics and AI*, 2:28, 2015.

[35] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 3169–3176, Washington, DC, USA, 2011. IEEE Computer Society.

[36] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1):60–79, 2013.

[37] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proceedings of the 2013 IEEE International Conference on Computer Vision*, ICCV '13, pages 3551–3558, Washington, DC, USA, 2013. IEEE Computer Society.

[38] M. Wang, S. Abdelfattah, N. Moustafa, and J. Hu. Deep gaussian mixture-hidden markov model for classification of eeg signals. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(4):278–287, Aug 2018.