Weakly Supervised Learning of Heterogeneous Concepts in Videos

Sohil Shah[1]([✉]), Kuldeep Kulkarni[2], Arijit Biswas[3], Ankit Gandhi[4],
Om Deshmukh[4], and Larry S. Davis[1]

[1] University of Maryland, College Park, MD, USA
sohilas@umd.edu
[2] Arizona State University, Tempe, AZ, USA
[3] Amazon Development Center India, Bangalore, India
[4] Xerox Research Centre India, Bangalore, India

Abstract. Typical textual descriptions that accompany online videos are 'weak': i.e., they mention the important heterogeneous concepts in the video but not their corresponding spatio-temporal locations. However, certain location constraints on these concepts can be inferred from the description. The goal of this paper is to present a generalization of the Indian Buffet Process (IBP) that can (a) systematically incorporate heterogeneous concepts in an integrated framework, and (b) enforce location constraints, for efficient classification and localization of the concepts in the videos. Finally, we develop posterior inference for the proposed formulation using mean-field variational approximation. Comparative evaluations on the Casablanca and the A2D datasets show that the proposed approach significantly outperforms other state-of-the-art techniques: 24 % relative improvement for pairwise concept classification in the Casablanca dataset and 9 % relative improvement for localization in the A2D dataset as compared to the most competitive baseline.

1 Introduction

Watching and sharing videos on social media has become an integral part of everyday life. We are often intrigued by the textual description of the videos and attempt to fast-forward to the segments of interest without watching the entire video. However, these textual descriptors usually do not specify the exact segment of the video associated with a particular description. For example, someone describing a movie clip as "head-on collision between cars while Chris Cooper is driving" neither provide the time-stamps for the collision or driving events nor the spatial locations of the cars or Chris Cooper. Such descriptions are referred to as 'weak labels'. For efficient video navigation and consumption, it is

---

important to automatically determine the spatio-temporal locations of these concepts (such as 'collision' or 'cars'). However, it is prohibitively expensive to train concept-specific models for all concepts of interest in advance and use them for localization. This shortcoming has triggered a great amount of interest in *jointly* learning concept-specific classification models as well as localizing concepts from multiple weakly labeled images [1–3] or videos [4,5].

Video descriptions include concepts which may refer to persons, objects, scenes and/or actions and thus a typical description is a combination of heterogeneous concepts. In the running example, extracted heterogeneous concepts are 'car' (object), 'head-on collision' (action), 'Chris Cooper' (person) and 'driving' (action). Learning classifiers for these heterogeneous concepts along with localization is an extremely challenging task because: (a) the classifiers for different kinds of concepts are required to be learned simultaneously, e.g., a face classifier, an object classifier, an action classifier etc., and (b) the learning model must take into account the spatio-temporal location constraints imposed by the descriptions while learning these classifiers. For example, the concepts 'head-on collision' and 'cars' should spatio-temporally co-occur at least once and there should be at least one car in the video.

Recently there has been growing interest to jointly learn concept classifiers from weak labels [1,5]. Bojanowski *et al.* [5] proposed a discriminative clustering framework to jointly learn person and action models from movies using weak supervision provided by the movie scripts. Since weak labels are extracted from scripts, each label can be associated with a particular shot in the movie, which may last only for a few seconds, i.e., the labels are well localized and that makes the overall learning easier. However, in real world videos, one does not have access to such shot-level labels but only to video-level labels. Therefore in our work, we do not assume availability of such well localized labels, and tackle the more general problem of learning concepts from weaker video-level labels. The framework in [5], when extended to long videos, does not give satisfactory results (see Sect. 4). Such techniques, which are based on a linear mapping from features to labels and model background using only a single latent factor, are usually inadequate to capture all the inter-class and intra-class variations. Shi *et al.* [1] jointly learn object and attribute classifiers from images using weakly supervised Indian Buffet Process (IBP). Note that IBP [6,7] allows observed features to be explained by a countably infinite number of latent factors. However, the framework in [1] is not designed to handle heterogeneous concepts and location constraints, which leads to a significant degradation in performance (Sect. 4.3). [8] and [9] propose IBP based cross-modal categorization/query image retrieval models which learn semantically meaningful abstract features from multimodal (image, speech and text) data. However, these unsupervised approaches do not incorporate any location constraints which naturally arise in the weakly supervised setting with heterogeneous labels.

We propose a novel Bayesian Non-parametric (BNP) approach called WSC-SIIBP (Weakly Supervised, Constrained &Stacked Integrative IBP) to jointly

learn heterogeneous concept classifiers and localize these concepts in videos. BNP models are a class of Bayesian models where the hidden structure that may have generated the observed data is not assumed to be fixed. Instead, a framework is provided that allows the complexity of the model to increase as more data is observed [10]. Specifically, we propose:

1. A novel generalization of IBP which for the first time incorporates weakly supervised spatio-temporal location constraints and heterogeneous concepts in an integrated framework.
2. Posterior inference of WSC-SIIBP model using mean-field approximation.

We assume that the weak video labels come in the form of tuples: in the running example, the extracted heterogeneous concept tuples are ({car, head-on collision}, {Chris Cooper, driving})[1]. We perform experiments on two video datasets (a) the Casablanca movie dataset [5] and (b) the A2D dataset [11]. We show that the proposed approach WSC-SIIBP outperforms several state-of-the-art methods for heterogeneous concept classification and localization in a weakly supervised setting. For example, WSC-SIIBP leads to a relative improvement of 7 %, 5 % and 24 % on person, action and pairwise classification accuracies, respectively, over the most competitive baselines on the Casablanca dataset. Similarly, the relative improvement on localization accuracy is 9 % over the next best approach on the A2D dataset.
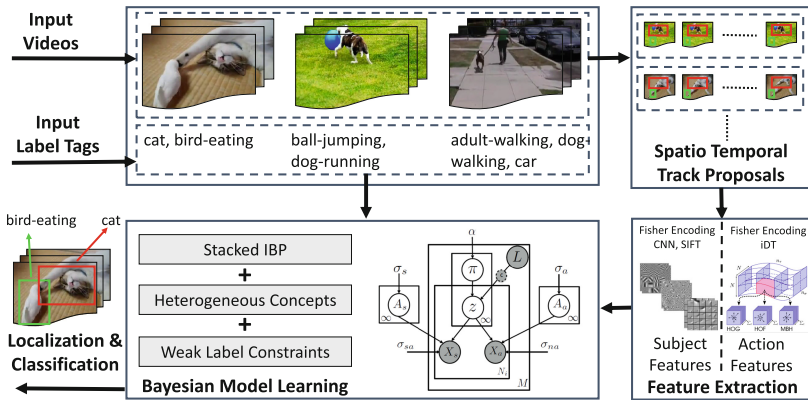

Fig. 1. Pipeline of WSC-SIIBP. Multiple videos with heterogeneous weak labels are provided as input and localization and classification of the concepts are performed in these videos.

---

[1] Extracting concept tuples is an interesting research problem but beyond the scope of this paper.

2   Related Work

In this section, we discuss relevant prior work in two broad categories.

Weakly Supervised Learning: Localizing concepts and learning classifiers from weakly annotated data is an active research topic. Researchers have learned models for various concepts from weakly labeled videos using Multi-Instance Learning (MIL) [12,13] for human action recognition [14], visual tracking [15] etc. Cour *et al.* [16] uses a novel convex formulation to learn face classifiers from movies and TV series using multimodal features which are obtained from finely aligned screenplay, speech and video data. In [4,17], the authors propose discriminative clustering approaches for aligning videos with temporally ordered text descriptions or predefined tags and in the process also learn action classifiers. In our approach, we consider weak labels which are neither ordered nor aligned to any specific video segment. [18] proposes a method for learning object class detectors from real world web videos known to contain only the target class by formulating the problem as a domain adaptation task. [19] learns weakly supervised object/action classifiers using a latent-SVM formulation where the objects or actions are localized in training images/videos using latent variables. We note that - both [18,19] consider only a single weak label per video and, unlike our approach, do not jointly learn the heterogeneous concepts. The authors in [20,21] use dialogues, scene and character identification to find an optimal mapping between a book and movie shots using shortest path or CRF approach. However, these approaches neither jointly model heterogeneous concepts nor spatio-temporally localized them. Although [22] proposes a discriminative clustering model for coreference resolution in videos, only faces are considered in their experiments.

Heterogeneous concept learning: There are prior works on automatic image [23–26] and video [27–29] caption generation, where models are trained on pairs of image/video and text that contain heterogeneous concept descriptions to predict captions for novel images/videos. While most of these approaches rely on deep learning methods to learn a mapping between an image/video and the corresponding text description, [25] uses MIL to learn visual concept detectors (spatial localization in images) for nouns, verbs and adjectives. However, none of these approaches spatio-temporally localize points of interests in videos. Perhaps the available video datasets are not large enough to train such a weakly supervised deep learning model.

To the best of our knowledge there is no prior work that jointly classifies and localizes heterogeneous concepts in weakly supervised videos.

3   WSC-SIIBP: Model and Algorithm

In this section, we describe the details of WSC-SIIBP (see Fig. 1 for the pipeline). We first introduce notations and motivate our approach in Sects. 3.1 and 3.2 respectively. This is followed by Sect. 3.3 where we introduce stacked nonparametric graphical model - IBP and its corresponding posterior computation.

In Sects. 3.4 and 3.5, we formulate an extension of the stacked IBP model which can generalize to heterogeneous concepts as well as incorporate the constraints obtained from weak labels. In Sect. 3.6, we briefly describe the inference procedure using truncated mean-field variational approximation and summarize our entire algorithm. Finally, we discuss how one can classify and localize concepts in new test videos using WSC-SIIBP.

3.1   Notation

Assume we are given a set of weakly labeled videos denoted by $\boldsymbol{\Lambda} = \{(i, \Gamma^{(i)})\}$, where i indicates a video and $\Gamma^{(i)}$ denotes the heterogeneous weak labels corresponding to the i-th video. Although the proposed approach can be used for any number of heterogeneous concepts, for readability, we restrict ourselves to two concepts and call them subjects and actions. We also have a closed set of class labels for these heterogeneous concepts: for subjects $\mathcal{S} = (s_1, \ldots, s_{K_s})$ and for actions $\mathcal{A} = (a_1, \ldots, a_{K_a})$. Let $K_s = |\mathcal{S}|$, $K_a = |\mathcal{A}|$, $\Gamma^{(i)} = \{(s_l, a_l) : s_l \in \mathcal{S} \cup \emptyset, a_l \in \mathcal{A} \cup \emptyset, 1 \le l \le |\Gamma^{(i)}|\}$, $\emptyset$ indicate that the corresponding subject or action class label is not present and $M = |\boldsymbol{\Lambda}|$ represents the number of videos. The video-level annotation simply indicates that the paired concepts $\Gamma^{(i)}$ can occur anywhere in the video and at multiple locations.

Assume that N_i spatio-temporal tracks are extracted from each video i where each track j is represented as an aggregation of multiple local features, $\mathbf{x}_j^{(i)}$. The spatio-temporal tracks could be face tracks, 3-D object proposals or action proposals (see Sect. 4.1 for more details). We associate the jth track in video i to an infinite binary latent coefficient vector $\mathbf{z}_j^{(i)}$ [1,6]. Each video i is represented by a bag of spatio-temporal tracks $\mathbf{X}^{(i)} = \{\mathbf{x}_j^{(i)}, j = 1, \ldots, N_i\}$. Similarly, $\mathbf{Z}^{(i)} = \{\mathbf{z}_j^{(i)}, j = 1, \ldots, N_i\}$.

3.2   Motivation

Our objective is to learn (a) a mapping between each of the N_i tracks in video i and the labels in $\Gamma^{(i)}$ and (b) the appearance model for each label identity such that the tracks from new test videos can be classified. To achieve these objectives, it is important for any model to discover the latent factors that can explain similar tracks across a set of videos with a particular label. In general, the number of latent factors are not known apriori and must be inferred from the data. In Bayesian framework, IBP treats this number as a random variable that can grow with new observations, thus letting the model to effectively explain the unbounded complexity in the data. Specifically, IBP defines a prior distribution over an equivalence class of binary matrices of bounded rows (indicating spatio-temporal tracks) and infinite columns (indicating latent coefficients). To achieve our goals, we build on IBP and introduce the WSC-SIIBP model which can effectively learn the latent factors corresponding to each heterogeneous concept and utilize prior location constraints to reduce the ambiguity in learning through the knowledge of other latent coefficients.

3.3   Indian Buffet Process (IBP)

The spatio-temporal tracks in the videos $\mathbf{\Lambda}$ are obtained from an underlying generative process. Specifically, we consider a stacked IBP model [1] as described below.

– For each latent factor $k \in 1 \ldots \infty$,
  1. Draw an appearance distribution with mean $\mathbf{a}_k \sim \mathcal{N}(0, \sigma_A^2 \mathbf{I})$
– For each video $i \in 1 \ldots M$,
  1. Draw a sequence of i.i.d. random variables, $v_1^{(i)}, v_2^{(i)} \cdots \sim \text{Beta}(\alpha, 1)$
  2. Construct the prior on the latent factors, $\pi_k^{(i)} = \prod_{t=1}^{k} v_t^{(i)}, \forall k \in 1 \ldots \infty$,
  3. For j^{th} subject track in i^{th} video, where $j \in 1 \ldots N_i$,
    (a) Sample state of each latent factor, $z_{jk}^{(i)} \sim \text{Bern}(\pi_k^{(i)})$,
    (b) Sample track appearance, $\mathbf{x}_j^{(i)} \sim \mathcal{N}\left(\mathbf{z}_j^{(i)} \mathbf{A}, \sigma_n^2 \mathbf{I}\right)$

where α is the prior controlling the sparsity of latent factors, σ_A^2 and σ_n^2 are the prior appearance and noise variance shared across all factors, respectively. Each $\mathbf{a}_k$ forms k^{th} row of $\mathbf{A}$ and the value of the latent coefficient $z_{jk}^{(i)}$ indicates whether data $\mathbf{x}_j^{(i)}$ contains the kth latent factor or not. In the above model, we have used stick-breaking construction [30] to generate the $\pi_k^{(i)}$s.

Posterior: Now, we describe how the posterior is obtained for the above graphical model. Let $\mathbf{Y} = \left\{\pi^{(1)} \ldots \pi^{(M)}, \mathbf{Z}^{(1)} \ldots \mathbf{Z}^{(M)}, \mathbf{A}\right\}$ and $\mathbf{\Theta} = \left\{\alpha, \sigma_A^2, \sigma_n^2\right\}$ denote hidden variables and prior parameters, respectively. $\mathbf{X}$ denotes the concatenation of all the spatio-temporal tracks in all M videos, $\left\{\mathbf{X}^{(1)} \ldots \mathbf{X}^{(M)}\right\}$. Given prior distribution $\Psi(\mathbf{Y}|\mathbf{\Theta})$ and likelihood function $p(\mathbf{x}_j^{(i)}|\mathbf{Y}, \mathbf{\Theta})$, the posterior probability is given by,

$$
\begin{aligned}
p(\mathbf{Y}|\mathbf{X}, \mathbf{\Theta}) &= \frac{\Psi(\mathbf{Y}|\mathbf{\Theta}) \prod_{i=1}^{M} \prod_{j=1}^{N_i} p(\mathbf{X}_j^{(i)}|\mathbf{Y}, \mathbf{\Theta})}{p(\mathbf{X}|\mathbf{\Theta})} \\
\Psi(\mathbf{Y}|\mathbf{\Theta}) &= \prod_{k=1}^{\infty} \left(\prod_{i=1}^{M} p(\pi_k^{(i)}|\alpha) \prod_{j=1}^{N_i} p(z_{jk}^{(i)}|\pi_k^{(i)})\right) p(\mathbf{a}_{k.}|\sigma_A^2).
\end{aligned}
\tag{1}
$$

where $p(\mathbf{X}|\mathbf{\Theta})$ is the marginal likelihood. For simplicity, we denote $p(\mathbf{Y}|\mathbf{X}, \mathbf{\Theta})$ as $q(\mathbf{Y})$. Apart from the significance of inferring $\mathbf{Z}^{(i)}$ for identifying track-level labels, inferring prior $\pi_k^{(i)}$ for each video helps to identify video-level labels, while the inference of appearance model $\mathbf{A}$ will be used to classify new test samples (see Sect. 3.6). Thus, learning in our model requires computing the full posterior distribution over $\mathbf{Y}$.

Regularized posterior: We note that it is difficult to infer the regularized posterior distributions using (1). However, it is known [31,32] that the posterior distribution in (1) can also be obtained as the solution $q(\mathbf{Y})$ of the following optimization problem,

$$
\min_{q(\mathbf{Y})} \quad \text{KL}\left(q(\mathbf{Y})||\Psi(\mathbf{Y}|\mathbf{\Theta})\right) - \sum_{i=1}^{M} \sum_{j=1}^{N_i} \int \log p(\mathbf{x}_j^{(i)}|\mathbf{Y}, \mathbf{\Theta}) q(\mathbf{Y}) d\mathbf{Y} \quad s.t. \quad q(\mathbf{Y}) \in P_{prob}
\tag{2}
$$

where KL(.) denotes the Kullback-Liebler divergence and P_{prob} is the probability simplex. As we will see later, this procedure enables us to learn the posterior distribution using a constrained optimization framework.

3.4 Integrative IBP

Our objective is to model heterogeneous concepts (such as subjects and actions) using a graphical model. However, the IBP model described above can not handle multiple concepts because it is highly unlikely that the subject and the action features can be explained by the same statistical model. Hence, we propose an extension of stacked IBP for heterogeneous concepts, where different concept types are modeled using different appearance models.

Let the subject and action types corresponding to the spatio-temporal track j in video i be denoted by $\mathbf{x^s}_j^{(i)}$ and $\mathbf{x^a}_j^{(i)}$, respectively, with each having different dimensions D^e ($e \in \{s, a\}$)[2]. Unlike the IBP model, $\mathbf{X^s}_j^{(i)}$ and $\mathbf{X^a}_j^{(i)}$ are now represented using two different gaussian noise models $\mathcal{N}(\mathbf{z}_j^{(i)}\mathbf{A}^s, \sigma_{ns}^2\mathbf{I})$ and $\mathcal{N}(\mathbf{z}_j^{(i)}\mathbf{A}^a, \sigma_{na}^2\mathbf{I})$ respectively where σ_{ne}^2 denotes prior noise variance and $\mathbf{A}^e$ are $K \times D^e$ matrices ($\mathrm{K} \to \infty$). The mean of the subject and action appearance models for each latent factor are also sampled independently from gaussian distributions of different variances σ_{Ae}^2. The new posterior probability is given by,

$$\tilde{q}(\mathbf{Y}) = \frac{\Psi(\mathbf{Y}|\mathbf{\Theta}) \prod_{i=1}^{M} \prod_{j=1}^{N_i} \prod_{e \in \{s,a\}} p(\mathbf{x^e}_j^{(i)}|\mathbf{Z}, \mathbf{A}^e, \mathbf{\Theta})}{p(\mathbf{X}|\mathbf{\Theta})}$$

$$\Psi(\mathbf{Y}|\mathbf{\Theta}) = \prod_{k=1}^{\infty} \left(\prod_{i=1}^{M} p(\pi_k^{(i)}|\alpha) \prod_{j=1}^{N_i} p(z_{jk}^{(i)}|\pi_k^{(i)}) \right) \prod_{e \in \{s,a\}} p(\mathbf{a}_k^e|\sigma_{Ae}^2\mathbf{I}). \tag{3}$$

3.5 Integrative IBP with Constraints

Although the graphical model described above is capable of handling heterogeneous features, the location constraints inferred from the weak labels still need to be incorporated into the graphical model. As motivated in Sect. 1, the concepts 'head-on collision' and 'cars' should spatio-temporally co-occur at least once and there should be at least one car in the full video. Imposing these location constraints in the inference algorithm can lead to more accurate parameter estimation of the graphical model and faster convergence of the inference procedure. These constraints can be generalized as follows,

1. Every label tuple in $\Gamma^{(i)}$, is associated with at least one spatio-temporal track (i.e., the event occurs in the video).
2. Spatio-temporal tracks should be assigned a label only from the list of weak labels assigned to the video. Concepts present in the video but not in the label will be subsumed in the background models.

---

[2] We often use e as a replacement for s and a throughout the paper.

Ideally, in the case of noiseless labels, these constraints should be strictly followed. However, we assume that real-world labels could be noisy and noise is independent of the videos. Hence, we allow constraints to be violated but penalize the violations using additional slack variables.

We associate the first K_s and the following K_a latent factors (the rows of $\mathbf{A}$) to the subject and action classes in $\mathcal{S}$ and $\mathcal{A}$ respectively. The inferred values of their corresponding latent coefficients in $\mathbf{z}_j^{(i)}$ are used to determine the presence/absence of the associated concept in a particular spatio-temporal track. The remaining unbounded number of latent factors are used to explain away the background tracks from unknown action and subject classes in a video. With these assignments, we enforce the following constraints on latent factors which are sufficient to satisfy the conditions mentioned earlier.

To satisfy 1, we introduce the following constraints, $\forall i \in 1 \dots M$, and $\forall j \in 1 \dots N_i$,

$$\sum_{j=1}^{N_i} z_{js}^{(i)} z_{ja}^{(i)} \geq 1 - \xi_{(s,a)}^{(i)}, \quad \forall (s,a) \in \Gamma^{(i)}, \tag{4}$$

$$\sum_{j=1}^{N_i} z_{js}^{(i)} \geq 1 - \xi_{(s,\emptyset)}^{(i)}, \quad \forall (s,\emptyset) \in \Gamma^{(i)}, \tag{5}$$

$$\sum_{j=1}^{N_i} z_{ja}^{(i)} \geq 1 - \xi_{(\emptyset,a)}^{(i)}, \quad \forall (\emptyset,a) \in \Gamma^{(i)}, \tag{6}$$

where ξ is the slack variable, z_{js} and z_{ja} are the latent factor coefficients corresponding to subject class s and action class a respectively.

To satisfy 2, we use the following constraints, $\forall i \in 1 \dots M$ and $\forall j \in 1 \dots N_i$,

$$z_{js}^{(i)} = 0, \text{if } \nexists (s,\emptyset) \in \Gamma^{(i)} \text{ and } \nexists (s,a) \in \Gamma^{(i)}, \forall a \in \mathcal{A}, \tag{7}$$

$$z_{ja}^{(i)} = 0, \text{if } \nexists (\emptyset,a) \in \Gamma^{(i)} \text{ and } \nexists (s,a) \in \Gamma^{(i)}, \forall s \in \mathcal{S}. \tag{8}$$

The constraints defined in (4)–(8) have been used in the context of discriminative clustering [5,22]. However, our model is the first to use these constraints in a Bayesian setup. In their simplest form, they can be enforced using the point estimate of z e.g., MAP estimation. However, $\mathbf{Z}^{(i)}$ is defined over the entire probability space. To enforce the above constraints in a Bayesian framework, we need to account for the uncertainty in $\mathbf{Z}^{(i)}$. Following [33,34], we define effective constraints as an expectation of the original constraints in (4)–(8), where the expectation is computed w.r.t. the posterior distribution in (3) (see supplementary material for the expectation constraints). The proposed graphical model, incorporating heterogeneous concepts as well as the location constraints provided by the weak labels, is shown in Fig. 2.

We restrict the search space for the posterior distribution in Eq. (3) by using the expectation constraints. In order to obtain the regularized posterior distribution of the proposed model, we solve the following optimization problem under these expectation constraints,
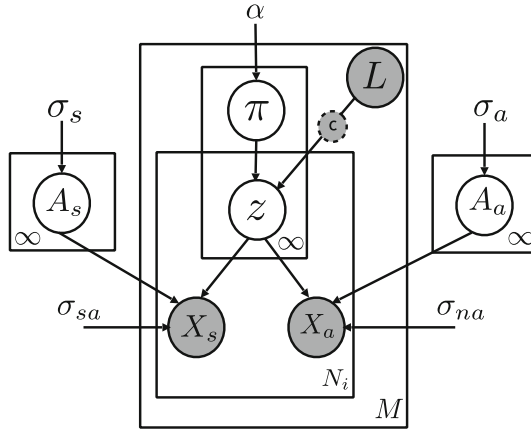
Fig. 2. WSC-SIIBP: Graphical Model using two heterogeneous concepts, subjects and actions. Each video (described by video-level labels L) is independently modeled using latent factor prior π and contains N_i tracks. Each track is represented using subject and action features X_s and X_a respectively, which are modeled using Gaussian appearance models A_s and A_a. z are the binary latent variables indicating the presence or absence of the latent factors in each track. c denotes the set of location constraints extracted from the video labels.

$$
\min_{\tilde{q}(\mathbf{Y}),\xi^{(i)}} \mathrm{KL}\left(\tilde{q}(\mathbf{Y})||\tilde{\Psi}(\mathbf{Y}|\mathbf{\Theta})\right) - \sum_{i=1}^{M}\sum_{j=1}^{N_i}\int\left(\sum_{e\in\{s,a\}}\log p\left(\mathbf{X}_j^{\mathbf{e}(i)}|\mathbf{Y},\mathbf{\Theta}\right)\right)\tilde{q}(\mathbf{Y})d\mathbf{Y}
$$
$$
+ C\sum_{i=1}^{M}\sum_{J\in\Gamma^{(i)}}\xi_J^{(i)} \quad s.t. \quad \tilde{q}(\mathbf{Y})\in P_{prob}
\tag{9}
$$

3.6 Learning and Inference

Note that the variational inference for true posterior $\tilde{q}(\mathbf{Y})$ (in Eq. (3)) is intractable over the general space of probability functions. To make our problem easier to solve, we establish *truncated* mean-field variational approximation [30] to the desired posterior $\tilde{q}(\mathbf{Y})$, such that the search space P_{prob} is constrained by the following tractable parametrised family of distributions,

$$
\tilde{w}(\mathbf{Y}) = \prod_{i=1}^{M}\left(\prod_{k=1}^{K_{max}}p(v_k^{(i)}|\tau_{k1}^{(i)},\tau_{k2}^{(i)})\prod_{j=1}^{N_i}p(z_{jk}^{(i)}|\nu_{jk}^{(i)})\right)\prod_{k=1}^{K_{max}}\prod_{e\in\{s,a\}}p(\mathbf{a}_k^e|\mathbf{\Phi}_k^e,\sigma_{ke}^2\mathbf{I}).
\tag{10}
$$

where $p(v_k^{(i)}|\tau_{k1}^{(i)},\tau_{k2}^{(i)}) = \mathrm{Beta}(v_k^{(i)};\tau_{k1}^{(i)},\tau_{k2}^{(i)})$, $p(z_{jk}^{(i)}|\nu_{jk}^{(i)}) = \mathrm{Bern}(z_{jk}^{(i)};\nu_{jk}^{(i)})$ and $p(\mathbf{a}_k^e|\mathbf{\Phi}_k^e,\sigma_{ke}^2\mathbf{I}) = \mathcal{N}(\mathbf{a}_k^e;\mathbf{\Phi}_k^e,\sigma_{ke}^2\mathbf{I})$. In Eq. (10), we note that all the latent variables are modeled independently of all other variables, hence simplifying the

inference procedure. The truncated stick breaking process of $\pi_k^{(i)}$'s is bounded at K_{max}, wherein $\pi_k = 0$ for $k > K_{max} \gg K_s + K_a + K_{bg}$. K_{bg} indicates the number of latent factors chosen to explain background tracks.

The optimization problem in Eq. (9) is solved using the posterior distribution from Eq. (10). We obtain the parameters (see supplementary material for details) $\sigma_{ke}^2, \Phi^{\mathbf{e}}{}_k, \tau_{ke}^{(i)}$ and $\nu_{jk}^{(i)}$ for the optimal posterior distribution $\tilde{q}(\mathbf{Y})$ using iterative update rules as summarized in Algorithm 1. We note that this algorithm is similar to other IBP learning algorithms [1,30]. The complexity of Algorithm 1 is $\mathcal{O}(MN_{max}D_{max}K_{max})$, the same as [1]. The mean of binary latent coefficients z_{jk}, denoted by ν_{jk}, has an update rule which will lead to several interesting observations.

$$\nu_{jk}^{(i)} = \frac{L_k^{(i)}}{1 + e^{-\zeta_{jk}^{(i)}}}. \tag{11}$$

$$\zeta_{jk}^{(i)} = \sum_{j=1}^{k} \left(\Psi(\tau_{j1}^{(i)}) - \Psi(\tau_{j1}^{(i)} + \tau_{j2}^{(i)}) \right) - \mathcal{L}_k - \sum_{e \in \{s,a\}} \frac{1}{2\sigma_{ne}^2} \left(D^e \sigma_{ke}^2 + \Phi^{\mathbf{e}}{}_k \Phi^{\mathbf{e}T}{}_k \right)$$

$$+ \sum_{e \in \{s,a\}} \frac{1}{\sigma_{ne}^2} \Phi^{\mathbf{e}}{}_k \left(\mathbf{x}_j^{(i)} - \sum_{l \neq k} \nu_{jl}^{(i)} \Phi^{\mathbf{e}}{}_l \right)^T + C \underbrace{\sum_{\substack{J \in \Gamma^{(i)} \\ J=(k,a)}} \mathbb{I}_{\left\{ \sum_{l=1}^{N_i} \nu_{lk}^{(i)} \nu_{la}^{(i)} < 1 \right\}} \nu_{ja}^{(i)}}_{(i)} \tag{12}$$

$$+ C \overbrace{\sum_{\substack{J \in \Gamma^{(i)} \\ J=(s,k)}} \mathbb{I}_{\left\{ \sum_{l=1}^{N_i} \nu_{ls}^{(i)} \nu_{lk}^{(i)} < 1 \right\}} \nu_{js}^{(i)}}^{(ii)} + C \overbrace{\mathbb{I}_{\left\{ \sum_{l=1}^{N_i} \nu_{lk}^{(i)} < 1, k \leq K_a + K_s \right\}}}^{(iii)}.$$

where $\Psi(.)$ is the digamma function, $\mathbb{I}$ is an indicator function, $L_k^{(i)}$ is an indicator variable and $\mathcal{L}_k$ is a lower bound for $\mathbb{E}_{\tilde{w}}[\log(1 - \prod_{j=1}^{k} v^{(i)})]$. The $L_k^{(i)}$ indicates whether a concept (action/subject) k is part of the ith video label set $\Gamma^{(i)}$ or not. If $L_k^{(i)} = 0$, all the corresponding binary latent coefficients $z_{jk}^{(i)}, j = \{1, \ldots, N_i\}$, are forced to 0, which is equivalent to enforcing the constraints in Eq. (7) and (8). Note that the value of $\nu_{jk}^{(i)}$ increases with $\zeta_{jk}^{(i)}$. The terms (i)-(iii) in the update rule for $\zeta_{jk}^{(i)}$ (Eq. (12)), which are obtained due to the location constraints in Eq. (4)–(6), act as the coupling terms between $\nu_{je}^{(i)}$'s. For example, for any action concept, term (ii) suggests that if the location constraints are not satisfied, better localization of all the coupled subject concepts (high value of $\nu_{js}^{(i)}$) will drive up the value of $\zeta_{ja}^{(i)}$. This implies that the strong localization of one concept can lead to better localization of other concepts.

The hyperparameter σ_{ne}^2 and σ_{Ae}^2 can be set apriori or estimated from data. Similar to the maximization step of EM algorithm, their empirical estimation can easily be obtained by maximizing the expected log-likelihood (see supplementary material).

Algorithm 1. Learning Algorithm of WSC-SIIBP

1: **Input:** data $\mathbf{\Lambda} = \{(i, \mathbf{\Gamma}^{(i)})\}_{i \in 1 \ldots M}$, constant α, K_{max}, C
2: **Output:** distribution $p(\mathbf{v}), p(\mathbf{Z}), p(\mathbf{A}^s)$ and $p(\mathbf{A}^a)$ and hyper-parameters $\sigma_{ns}^2, \sigma_{na}^2, \sigma_{As}^2$ and σ_{Aa}^2
3: **Initialize:** $\tau_{k1}^{(i)} = \alpha, \tau_{k2}^{(i)} = 1, \nu_{jk}^{(i)} = 0.5, \mathbf{\Phi}_k^s = \mathbf{\Phi}_k^a = 0, \sigma_{ks}^2 = \sigma_{ka}^2 = \sigma_{ns}^2 = \sigma_{na}^2 = \sigma_{As}^2 = \sigma_{Aa}^2 = 1$
4: **repeat**
5:      **repeat**
6:          update σ_{ke}^2 and $\mathbf{\Phi^e}_{k.}$, $\forall 1 \leq k \leq K_{max}, e \in \{s, a\}$;
7:          update $\tau_{k1}^{(i)}$ and $\tau_{k2}^{(i)}$, $\forall 1 \leq k \leq K_{max}$ and $i \in 1$ to M;
8:          update $\nu_{jk}^{(i)}$ using Equation (11) and (12), $\forall 1 \leq k \leq K_{max}, 1 \leq j \leq N_i$ and $i \in 1$ to M;
9:      **until** T iterations or $\frac{\|L(t-1)-L(t)\|}{L(t)} \leq 1e^{-3}$
10:      update the hyperparameters $\sigma_{As}^2, \sigma_{Aa}^2, \sigma_{ns}^2, \sigma_{na}^2$
11: **until** T' iterations or $\frac{\|L(t'-1)-L(t')\|}{L(t')} \leq 1e^{-4}$

Given the input features $\mathbf{X_s}$ and $\mathbf{X_a}$, the inferred latent coefficients $\nu_{je}^{(i)}$ estimate presence/absence of associated classes in a video. One can classify each spatio-temporal track by estimating the track-level labels using $L_j^* = \arg\max_k \nu_{jk}$. Here the maximization is over the latent coefficients corresponding to either the subject or action concepts depending upon the label which we are interested in extracting. For the concept localization task in a video with label pair (s, a), the best track in the video is selected using $j^* = \arg\max_j \nu_{js} \times \nu_{ja}$.

Test Inference: Although the above formulation is proposed for concept classification and localization in a given set of videos (transductive setting), the same algorithm can also be applied to unseen test videos. The latent coefficients for the tracks of test videos can be learned alongside the training data except that the parameters σ_{ke}^2, $\mathbf{\Phi^e}_{k.}$, σ_{Ae}^2 and σ_{ne}^2 are updated only using training data. In the case of free annotation, i.e., absence of labels for test video i, we run the proposed approach by setting $L_k^{(i)} = 1$ in Eq. (11), indicating that the tracks in a video i can belong to any of the classes in $\mathcal{S}$ or $\mathcal{A}$ (i.e., no constraints as defined by (4)–(8) are enforced).

4   Experimental Results

In this section, we present an evaluation of WSC-SIIBP on two real-world databases: Casablanca movie and A2D dataset, which represent typical 'in-the-wild' videos with weak labels on heterogeneous concepts.

4.1   Datasets

Casablanca dataset: This dataset, introduced in [5], has 19 persons (movie actors) and three action classes (sitdown, walking, background). The heterogeneous concepts used in this dataset are persons and actions. The Casablanca movie is divided into shorter segments of duration either 60 or 120 s. We manually annotate all the tracks in each video segment which may contain multiple persons and actions. Given a video segment and the corresponding video-level

labels (extracted from all ground truth track labels), our algorithm maps each of these labels to one or more tracks in that segment, i.e., converts the weak labels to strong labels. Our main objective of evaluation on this dataset is to compare the performance of various algorithms in classifying tracks from videos of varying length.

For our setting, we consider face and action as the two heterogeneous concepts and thus it is required to extract the face and the corresponding action track features. We extract 1094 facial tracks from the full 102 min Casablanca video. The face tracks are extracted by running the multi-view face detector from [35] in every frame and associating detections across frames using point tracks [36]. We follow [37] to generate the face track feature representations: Dense rootSIFT features are extracted for each face in the track followed by PCA and video-level Fisher vector encoding. The action tracks corresponding to 1094 facial tracks are obtained by extrapolating the face bounding-boxes using linear transformation [5]. For action features, we compute Fisher vector encoding on dense trajectories [38] extracted from each action track.

On an average, each 60 s. segment contains 11 face-action tracks and 4 face-action annotations while each 120 s. video contains 21 tracks and 6 annotations. Note that, our experimental setup is more difficult compared to the experimental setting considered in [5]. In [5], the Casablanca movie is divided into numerous bags based on the movie script, where on average each segment is of duration 31 s. containing only 6.27 face-action tracks.

A2D dataset: This dataset [11] contains 3782 YouTube videos (on average 7–10 s. long) covering seven objects (bird, car etc.) performing one of nine actions (fly, jump etc.). The heterogeneous concepts considered are objects and actions. This dataset provides the bounding box annotations for every video label pair of object and action. Using the A2D dataset, we aim to analyze the track localization performance on weakly labeled videos as well as the track classification accuracy on a held-out test dataset.

We use the method proposed in [39] to generate spatio-temporal object track proposals. For computational purpose, we consider only 10 tracks per video and use the Imagenet pretrained VGG CNN-M network [40] to generate object feature representation. We extract convolutional layer conv-4 and conv-5 features for each track image followed by PCA and video-level Fisher vector encoding. In this dataset, the corresponding action tracks are kept similar to the object tracks (proposals) and the action features are extracted using the same approach as used for the Casablanca dataset.

4.2   Baselines

We compare WSC-SIIBP to several state-of-the-art approaches using the same features.

1. **WS-DC** [5]: This approach uses similar weak constraints as in (4)–(6), but in a discriminative setup where the constraints are incorporated in a biconvex optimization framework.
2. **WS-SIBP** [1]: This is a weakly supervised stacked IBP model which does not consider integrative framework for heterogeneous data and only enforces constraints equivalent to (7)–(8). For each spatio-temporal track, the features extracted for heterogeneous concepts are concatenated while using this approach.
3. **WS-S / WS-A**: This is similar to WS-SIBP except that instead of concatenating features from multiple concepts they are treated independently in two different IBP. WS-S(WS-A) is used to model only the person/object(action) features.
4. **WS-SIIBP**: This model integrates WS-SIBP with heterogeneous concepts.
5. **WSC-SIBP**: This model is similar to WS-SIBP, but unlike WS-SIBP, it additionally enforces the location constraints obtained from weak labels.

Implementation details: For each dataset, the Fisher encoded features are PCA reduced to an appropriate dimension, D^e. We select the best feature length and other algorithm specific hyper-parameters for each algorithm using cross-validation on a small set of input videos. For the IBP based models, the cross-validation range for hyper-parameters are $K_{max} := K_a + K_s : 10 : K_a + K_s + 100$, $\alpha := 3K_{max} : 10 : 4K_{max}$ and $C := 0 : 0.5 : 5$. For all IBP based models, the parameters D^e, α, K_{max} and C are set as 32, 100, 30 and 0.5 respectively for the Casablanca dataset and as 128, 160, 50 and 5 respectively for the A2D dataset. For WS-DC, D^e is set as 1024.

4.3 Results on Casablanca

The track-level classification performance is compared in Fig. 3. From Figs. 3a and d, it can be seen that WSC-SIIBP significantly outperforms other methods for person and action classification in almost all of the scenarios. For instance, in the 120 s video segments, person classification improves by 4 % (relative improvement is 7 %) compared to the most competitive approach WS-SIBP. We also compare pairwise label accuracy to gain insight into the importance of the constraints in eq (4)–(6). For any given track with non-background person and action label, the classification is assumed to be correct only if both person and action labels are correctly assigned. Even in this scenario WSC-SIIBP performs 8.1 % better (24 % relative improvement) than the most competitive baseline. Since we combine the heterogeneous concepts along with location constraints in an integrated framework, WSC-SIIBP outperforms all other baselines. The weak results of WS-DC in pairwise classification, though surprising, can be attributed to their action classification results which are significantly biased towards one particular action 'sitdown' (Fig. 3d, note that WS-DC performs very poorly in 'walking' classification). Indeed, it should be noted that nearly 40 % and 89 % of

person and action labels respectively belong to the background class. Thus, for fair evaluation of both background and non-background classes, we also plot the recall of background class against the recall of nonbackground classes for person and action classification in Fig. 3b, c, e and f. These curves were obtained by simultaneously computing recall for background and non-background classes over a range of threshold values on score, ν. The mean average precision (mAP) of WSC-SIIBP along with all other baselines are plotted in Fig. 3g and h. The mAP values also clearly demonstrate the effectiveness of the proposed approach. From the performance of WS-SIIBP (integrative concepts, no constraints) and WSC-SIBP (no integrative concepts, constraints) (Fig. 3a and d), it is clear that the improvement in performance in the WSC-SIIBP can be attributed to both addition of integrative concepts and the location constraints.

Effect of constraints (7)**,** (8)**:** We note that, regardless of other differences, every weakly supervised IBP model considered here enforces constraints (7), (8). However, these constraints are not part of the original WS-DC. To make a fair comparison between WS-DC and WSC-SIIBP, we analyze the effect of these constraints in Fig. 3i. Although, these additional constraints improve WS-DC performance, they do not supersede the performance of WSC-SIIBP. Further we
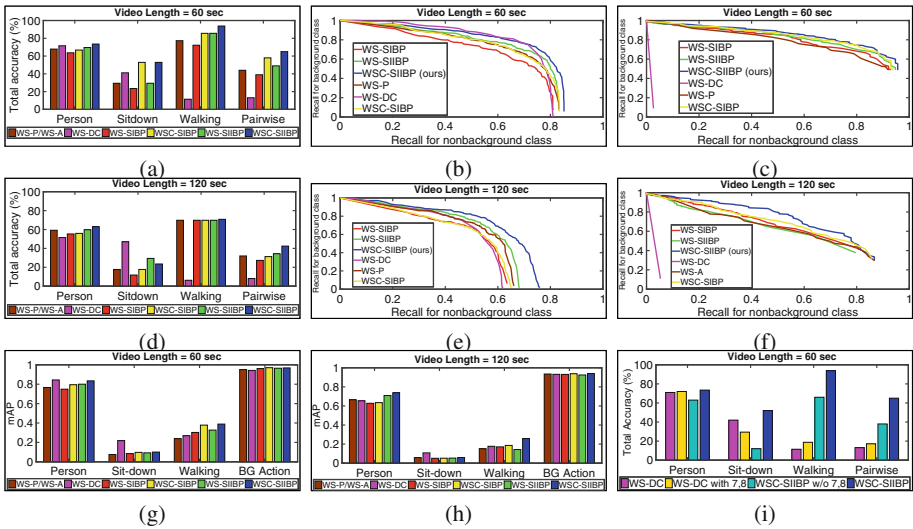

Fig. 3. Comparison of results for the Casablanca movie dataset. (a) Classification accuracy for 60 s. segments. (b) Recall for background vs non-background class (60 s, person). (c) Recall for background vs non-background (60 s., action). (d) Classification accuracy for 120 s. segments. (e) Recall for background vs non-background class (120 s., person). (f) Recall for background vs non-background (120 s., action). (g),(h) Mean Average Precision for 60, 120 s. segments. (i) Classification accuracy obtained with and without constraints (7) and (8)

observe that these constraints have improved the performance of all the weakly supervised IBP models.

4.4   Results on A2D

First, we evaluate localization performance on the full A2D dataset. We experiment with 37,820 tracks extracted from 3,782 videos with around 5000 weak labels. For every given object-action label pair our algorithm selects the best track from the corresponding video using the approach outlined in Sect. 3.6. The localization accuracy is measured by calculating the average IoU (Intersection over Union) of the selected track (3-D bounding box) with the ground truth bounding box. The class-wise IoU accuracy and the mean IoU accuracy for all classes are tabulated in Tables 1 and 2 respectively. In this task WSC-SIIBP also leads to a relative improvement of 9 % above the next best baseline. We also evaluate how accurately the extracted object proposals match with the ground truth bounding boxes to estimate an upper bound on the localization accuracy (referred as Upper Bound in Tables 1 and 2). In this case, the track maximizing the average IoU with the ground truth annotation is selected and the corresponding IoU is reported. We plot the correct localization accuracy with varying IoU thresholds in Fig. 4a, which also shows the effectiveness of the proposed approach. Figure 4b–4c shows some qualitative track localization results using the proposed approach on the selected frame.

Test Inference: We evaluate the classification performance on held-out test samples using the same train/test partition as in [11]. We consider two setups for the evaluation, (a) using video-level labels for the test samples and (b) free annotation where no test video labels are provided. The proposed approach is compared with GT-SVM, which is a fully supervised linear SVM that uses ground truth bounding boxes and their corresponding strong labels during training. The results are tabulated in Table 3. Note that the performance of WSC-SIIBP is close to that of the fully supervised setup.
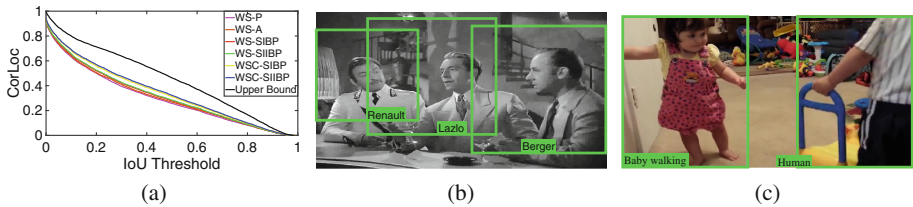

Fig. 4. (a) Correct localization accuracy at various IOU thresholds. (b) and (c) Qualitative results: green boxes show the concept localization using our proposed approach.

Table 1. Per class mean IoU on A2D dataset.

	adult	baby	ball	bird	car	cat	dog	climb	crawl	eat	fly	jump	roll	run	walk
WSC-SIIBP	28.4	43.6	9.8	37.8	37.4	40.8	42.0	37.5	47.6	46.1	24.5	29.4	50.9	25.6	37.2
Upper Bound	39.9	53.9	16.4	48.2	48.7	52.8	51.4	50.0	59.2	57.2	33.9	41.0	59.1	38.1	47.9

Table 2. Average IoU comparison with other approaches on A2D dataset.

	Random	WS-P	WS-A	WS-SIBP	WS-SIIBP	WSC-SIBP	WSC-SIIBP	Upper Bound
IoU	25.5	29.7	30.43	31.1	31.55	31.69	**34.38**	45.05

Table 3. mAP classification test accuracy on A2D dataset.

	WSC-SIIBP		GT-SVM	
Setup	Obj	Act	Obj	Act
Using video Labels	94.77	90.68	98.20	94.92
Free Annotation	76.62	64.77	85.18	73.26

5   Conclusion

We developed a Bayesian non-parametric approach that integrates the Indian Buffet Process with heterogeneous concepts and spatio-temporal location constraints arising from weak labels. We report experimental results on two recent datasets containing heterogeneous concepts such as persons, objects and actions and show that our approach outperforms the best state of the art method. In future work, we will extend the WSC-SIIBP model to additionally localize audio concepts from speech input and develop an end-to-end deep neural network for joint feature learning and Bayesian inference.

References

1. Shi, Z., Yang, Y., Hospedales, T.M., Xiang, T.: Weakly supervised learning of objects, attributes and their associations. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 472–487. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10605-2_31
2. Leung, T., Song, Y., Zhang, J.: Handling label noise in video classification via multiple instance learning. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2056–2063. IEEE (2011)
3. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free? weakly-supervised learning with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)

 4. Bojanowski, P., Lajugie, R., Bach, F., Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Weakly supervised action labeling in videos under ordering constraints. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 628–643. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10602-1_41
 5. Bojanowski, P., Bach, F., Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Finding actors and actions in movies. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 2280–2287. IEEE (2013)
 6. Ghahramani, Z., Griffiths, T.L.: Infinite latent feature models and the indian buffet process. Adv. Neural Inf. Proces. Syst. **18**, 475–482 (2005)
 7. Griffiths, T.L., Ghahramani, Z.: The indian buffet process: an introduction and review. J. Mach. Learn. Res. **12**, 1185–1224 (2011)
 8. Ozdemir, B., Davis, L.S.: A probabilistic framework for multimodal retrieval using integrative indian buffet process. In: Advances in Neural Information Processing Systems, pp. 2384–2392 (2014)
 9. Yildirim, I., Jacobs, R.A.: A rational analysis of the acquisition of multisensory representations. Cogn. Sci. **36**(2), 305–332 (2012)
10. Gershman, S.J., Blei, D.M.: A tutorial on bayesian nonparametric models. J. Math. Psychol. **56**(1), 1–12 (2012)
11. Xu, C., Hsieh, S.H., Xiong, C., Corso, J.J.: Can humans fly? action understanding with multiple classes of actors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2264–2273 (2015)
12. Zhang, C., Platt, J.C., Viola, P.A.: Multiple instance boosting for object detection. Adv. Neural Inf. Process. Syst. **7**, 1417–1424 (2005)
13. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Proceedings of Advances in Neural Information Processing Systems, pp. 561–568 (2002)
14. Ali, S., Shah, M.: Human action recognition in videos using kinematic features and multiple instance learning. Pattern Anal. Mach. Intell. IEEE Trans. **32**(2), 288–303 (2010)
15. Babenko, B., Yang, M.H., Belongie, S.: Visual tracking with online multiple instance learning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 983–990. IEEE (2009)
16. Cour, T., Jordan, C., Miltsakaki, E., Taskar, B.: Movie/Script: alignment and parsing of video and text transcription. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5305, pp. 158–171. Springer, Heidelberg (2008). doi:10.1007/978-3-540-88693-8_12
17. Bojanowski, P., Lagugie, R., Grave, E., Bach, F., Laptev, I., Ponce, J., Schmid, C.: Weakly-supervised alignment of video with text. In: ICCV, IEEE (2015)
18. Prest, A., Leistner, C., Civera, J., Schmid, C., Ferrari, V.: Learning object class detectors from weakly annotated video. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3282–3289. IEEE (2012)
19. Bilen, H., Namboodiri, V.P., Van Gool, L.J.: Object and action classification with latent window parameters. Int. J. Comput. Vis. **106**(3), 237–251 (2014)
20. Tapaswi, M., Bauml, M., Stiefelhagen, R.: Book2movie: Aligning video scenes with book chapters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1827–1835 (2015)
21. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning books and movies: towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 19–27 (2015)

22. Ramanathan, V., Joulin, A., Liang, P., Fei-Fei, L.: Linking people in videos with "their" names using coreference resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 95–110. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10590-1_7

23. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)

24. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of The 32nd International Conference on Machine Learning, pp. 2048–2057 (2015)

25. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al.: From captions to visual concepts and back. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1473–1482 (2015)

26. Sun, C., Gan, C., Nevatia, R.: Automatic concept discovery from parallel text and visual corpora. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2596–2604 (2015)

27. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence-video to text. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4534–4542 (2015)

28. Rohrbach, A., Rohrbach, M., Schiele, B.: The long-short story of movie description. In: Gall, J., Gehler, P., Leibe, B. (eds.) GCPR 2015. LNCS, vol. 9358, pp. 209–221. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24947-6_17

29. Cho, K., Courville, A., Bengio, Y.: Describing multimedia content using attention-based encoder-decoder networks. Multimedia IEEE Trans. **17**(11), 1875–1886 (2015)

30. Doshi, F., Miller, K., Gael, J.V., Teh, Y.W.: Variational inference for the indian buffet process. In: International Conference on Artificial Intelligence and Statistics, pp. 137–144 (2009)

31. Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational inference. Found. Trends Mach. Learn. **1**(1–2), 1–305 (2008)

32. Zellner, A.: Optimal information processing and bayes's theorem. Am. Stat. **42**(4), 278–280 (1988)

33. Zhu, J., Chen, N., Xing, E.P.: Bayesian inference with posterior regularization and applications to infinite latent svms. J. Mach. Learn. Res. **15**(1), 1799–1847 (2014)

34. Ganchev, K., Graça, J., Gillenwater, J., Taskar, B.: Posterior regularization for structured latent variable models. J. Mach. Learn. Res. **11**, 2001–2049 (2010)

35. Zhu, X., Ramanan, D.: Face detection, pose estimation and landmark estimation in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)

36. Everingham, M., Sivic, J., Zisserman, A.: Hello! my name is buffy-automatic naming of characters in tv video. In: BMVC. vol. 2. 6 (2006)

37. Parkhi, O.M., Simonyan, K., Vedaldi, A., Zisserman, A.: A compact and discriminative face track descriptor. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1693–1700. IEEE (2014)

38. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 3551–3558. IEEE (2013)
39. Oneata, D., Revaud, J., Verbeek, J., Schmid, C.: Spatio-temporal object detection proposals. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 737–752. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10578-9_48
40. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: British Machine Vision Conference (2014)