Fast Supervised LDA for Discovering Micro-Events in Large-Scale Video Datasets

Angelos Katharopoulos, Despoina Paschalidou, Christos Diou, Anastasios Delopoulos
Multimedia Understanding Group
ECE Department, Aristotle University of Thessaloniki, Greece
{katharas, pdespoin}@auth.gr; diou@mug.ee.auth.gr; adelo@eng.auth.gr

ABSTRACT

This paper introduces fsLDA, a fast variational inference method for supervised LDA, which overcomes the computational limitations of the original supervised LDA and enables its application in large-scale video datasets. In addition to its scalability, our method also overcomes the drawbacks of standard, unsupervised LDA for video, including its focus on dominant but often irrelevant video information (e.g. background, camera motion). As a result, experiments in the UCF11 and UCF101 datasets show that our method consistently outperforms unsupervised LDA in every metric. Furthermore, analysis shows that class-relevant topics of fsLDA lead to sparse video representations and encapsulate high-level information corresponding to parts of video events, which we denote "micro-events".

Keywords

video event detection; video micro-events; supervised topic modeling; variational inference

1. INTRODUCTION

Recently there have been significant advancements in video event detection with the two-stream models of [10, 15, 16] achieving high classification results. However, several related problems still exist (e.g. mapping motion to text descriptions, zero-shot detection) and seeking meaningful ways to analyse and represent video events remains a challenging and interesting problem. This paper aims to develop a method capable of decomposing actions or events from large-scale video data to a set of meaningful discriminative components, namely micro-events. To achieve that, we develop fsLDA, a variational inference algorithm that simultaneously preserves the computational efficiency of Latent Dirichlet Allocation (LDA) [2] and improves the discriminativeness of supervised LDA (sLDA) [1].

LDA was initially introduced as an unsupervised method, namely only the words in documents were considered to be the observed information. The goal was to infer topic distributions that maximized the likelihood of the data. Later on, Blei et al. [1] introduced a supervised variant of LDA in order to retain the discriminative information from the words with respect to the classes.

Supervised LDA has been successfully applied for classifying multimedia content mainly in the form of images [13, 8, 17]. Using a generative model is shown to be beneficial since it allows for different types of information to be encoded in a single latent space, resulting in improvements shared accross all predictions. For instance, modeling annotations in [13] marginally improves classification performance. However, it is shown in [3] that in sLDA the classes influence only lightly the latent topic assignments resulting in performance similar to LDA, which is an issue addressed in this paper. Furthermore, sLDA becomes computationally intractable for even moderately large video collections [9].

Regarding multimedia content, LDA has also been recently used, in a straightforward manner [9, 4], to combine information from different modalities and discover topics that span different vocabularies. To the best of our knowledge, there has been no attempt to train a discriminative LDA model on large-scale video data and this is probably due to computational issues in the traditional supervised LDA.

Unsupervised topic models encode the dominant structure in documents. In multimedia and especially in video data most content may refer to background movements and camera motion. This is less relevant to the depicted action than the foreground objects and their motion, thus our goal is to develop a method capable of encoding the most relevant information about the illustrated event rather than the most common.

The main contributions of this work are the following:

- We adopt topic modeling to infer both discriminative and semantic topics from large-scale video data, which encapsulate information about the micro-events that generate the events.
- We propose a supervised variation of LDA (fsLDA) that not only has lower asymptotic complexity than sLDA but is also able to adapt the influence of the supervised part on the final topic representation.

The rest of this paper is structured as follows. In section 2 the proposed method is presented. Experimental results are reported in section 3, followed by conclusions in section 4.

MM '16, October 15-19, 2016, Amsterdam, Netherlands
© 2016 ACM. ISBN 978-1-4503-3603-1/16/10...$15.00
DOI: http://dx.doi.org/10.1145/2964284.2967237

2. FAST SUPERVISED LDA

The key idea behind fsLDA is to reduce the computational complexity of sLDA so that it can infer topic distributions from large-scale video data, while at the same time increasing the influence the class information exerts on the topics to improve classification performance. Each video in the corpus corresponds to a mixture of visual and motion codewords, which are obtained by using a bag-of-words aggregation method on extracted local features. Every video belongs to one of C discrete classes, let it be y.

We suppose that the corpus consists of D documents and that there are V terms in the vocabulary. In addition, we denote two latent variables $\theta_{D\times K}$ and $z_{D\times N\times K}$, where K is the total number of topics and N is the number of codewords in a document d. The first hidden variable θ corresponds to the per-document topic distributions, while z corresponds to the per-codeword topic assignments. Moreover, α_K, $\beta_{K\times V}$ and $\eta_{K\times C}$ are the model parameters. sLDA assumes the following generative process for each document d in the corpus. For clarity, the subscript d is omitted when an equation refers to a single document (e.g. θ instead of θ_d and z_n instead of z_{dn}).

1. Draw topic proportions $\theta \sim \mathrm{Dir}(\alpha)$
2. For each codeword:
   (a) Draw topic assignment $z_n \mid \theta \sim \mathrm{Mult}(\theta)$
   (b) Draw word assignment $w_n \mid z_n, \beta_{1:K} \sim \mathrm{Mult}(\beta_{z_n})$
3. Draw class label $y \mid z_{1:N} \sim \mathrm{softmax}\left(\frac{1}{N}\sum_{n=1}^{N} z_n, \eta\right)$ where the softmax function provides the following distribution

$$p(y,\bar{z},\eta) = \frac{\exp\left(\eta_y^T \frac{1}{N}\sum_{n=1}^{N} z_n\right)}{\sum_{\hat{y}=1}^{C}\exp\left(\eta_{\hat{y}}^T \frac{1}{N}\sum_{n=1}^{N} z_n\right)}$$

Given a document and the corresponding class label the posterior distribution of the latent variables is intractable, thus we use variational methods to approximate it. Following Blei et al. [1], our goal is to maximize the evidence lower bound (ELBO) $\mathcal{L}(\cdot)$, which is given in equation 1.

$$\begin{aligned}&\log p(w, y \mid \alpha, \beta, \eta) \geq \mathcal{L}(\gamma, \phi \mid \alpha, \beta, \eta) = \\ &\mathbb{E}_q[\log p(\theta \mid \alpha)] + \mathbb{E}_q[\log p(z \mid \theta)] + \mathbb{E}_q[\log p(w \mid \beta, z)] + \\ &H(q) + \mathbb{E}_q[\log p(y \mid z, \eta)]\end{aligned} \tag{1}$$

The expectation is taken with respect to a variational distribution $q(\theta, z_{1:N} \mid \gamma, \phi_{1:N}) = q(\theta \mid \gamma)\prod_{n=1}^{N} q(z_n \mid \phi_n)$, where ϕ_n is the variational multinomial parameter for the topic assignment z_n and γ is the variational Dirichlet parameter. Figure 1 depicts the probabilistic graphical model of sLDA.

2.1 Inference approach of sLDA

In this section we present the inference approach of sLDA as it was introduced by Wang et al. in [13]. We point out the demerits of this approach, that motivated the development of our algorithm which is presented in section 2.2. In equation 1 the first three terms and the entropy $H(q)$ of the variational distribution are identical to the corresponding terms in the ELBO for unsupervised LDA [2]. The last term is the expected log probability of the class variable given the topic assignments, which can be computed from equation 2.

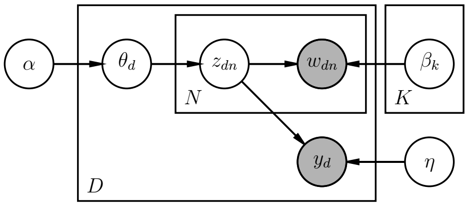

Figure 1: A graphical model representation of Supervised LDA

$$\mathbb{E}_q[\log p(y \mid z, \eta)] = \eta_y^T \sum_{n=1}^{N} \frac{\phi_n}{N} - \mathbb{E}_q\left[\log \sum_{\hat{y}=1}^{C} \exp(\eta_{\hat{y}}^T \bar{z})\right] \tag{2}$$

Following Wang et al. [13], the second term is approximated using Jensen's inequality, which results to the following update rules for ϕ and γ in the expectation step, where $\Psi(\cdot)$ is the first derivative of the $\log\Gamma(\cdot)$ and h is computed as in [13].

$$\begin{aligned}&\gamma = \alpha + \sum_{n=1}^{N}\phi_n \\ &\phi_n \propto \exp\left(\Psi(\gamma) + \frac{1}{N}\eta_y - (h^T\phi_n^{\mathrm{old}})^{-1}h\right)\end{aligned} \tag{3}$$

Notice that in order to update ϕ_n once, one must compute the rule multiple times, since it is a fixed-point iteration method. The supervised part of this update rule is multiplied by $\frac{1}{N}$ and as a result, the influence of the supervised part is reduced for documents with more words, also noticed in [3]. Another drawback of the traditional inference approach of sLDA concerns the maximization step, where one must keep in memory all the variational parameters in order to compute the gradient of $\mathcal{L}$ with respect to η (e.g. for a moderately large video collection with 5000 videos, 4000 codewords and 50 topics, the variational parameter ϕ consists of 10^9 elements).

2.2 Inference approach of fsLDA

The main goal of our method is to improve the computation of the approximate variational distribution and reduce the memory requirements. $\mathcal{L}$ in terms of ϕ_n is:

$$\begin{aligned}\mathcal{L}_{\phi_n} = &\sum_{i=1}^{K}\phi_{n,i}\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{K}\gamma_j)\right) + \\ &\sum_{i=1}^{K}\phi_{n,i}\log\beta_{i,n} - \sum_{i=1}^{K}\phi_{n,i}\log\phi_{n,i} + \\ &\eta_y^T\frac{\phi_n}{N} - \mathbb{E}_q\left[\log\sum_{\hat{y}=1}^{C}\exp(\eta_{\hat{y}}^T\bar{z})\right]\end{aligned} \tag{4}$$

The last term of equation 4 prevents a closed form solution for the update of ϕ_n. Using Jensen's inequality we can derive a lower bound for the expectation of the log normalizer and then approximate it with a second-order Taylor expansion, as shown in equation 5.

$$-\mathbb{E}_q\left[\log\sum_{\hat{y}=1}^{C}\exp(\eta_{\hat{y}}^T\bar{z})\right] \geq -\log\sum_{\hat{y}=1}^{C}\mathbb{E}_q\left[\exp(\eta_{\hat{y}}^T\bar{z})\right] \approx -\log\sum_{\hat{y}=1}^{C}\exp(\eta_{\hat{y}}^T\,\mathbb{E}_q[\bar{z}])\left(1+\frac{1}{2}\eta_{\hat{y}}^T\,\mathbb{V}_q[\bar{z}]\eta_{\hat{y}}\right) \tag{5}$$

The expectation is $\mathbb{E}_q[\bar{z}] = \frac{1}{N}\sum_{n=1}^{N}\phi_n$, while the variance is $\mathbb{V}_q[\bar{z}] = \frac{1}{N^2}\left(\sum_{n=1}^{N}\sum_{m\neq n}\phi_n\phi_m^T + \sum_{n=1}^{N}\mathrm{diag}(\phi_n)\right)$. It can be observed that the variance term is multiplied with a very small number, especially in the case of multimedia. For instance, in case of video data, the corresponding number of codewords usually surpasses 20,000. Therefore, we decide to approximate the expectation of the log normalizer with the first-order expansion.

Finding a closed form solution to maximize $\mathcal{L}_{\phi_n}$ requires computing the derivative with respect to ϕ_n and adding Lagrange Multipliers, $\mathcal{L}'_{\phi_n} = \mathcal{L}_{\phi_n} + \lambda_n\left(\sum_{i=1}^{K}\phi_{n,i} - 1\right)$.

$$\frac{\mathrm{d}\mathcal{L}'_{\phi_n}}{\mathrm{d}\phi_n} = \left(\Psi(\gamma) - \Psi(\sum_{j=1}^{K}\gamma_j)\right) + \log\beta_n - \log\phi_n - 1 + \lambda_n + \frac{1}{N}\left(\eta_y - \frac{\sum_{\hat{y}=1}^{C}\exp\left(\eta_{\hat{y}}^T\,\mathbb{E}_q[\bar{z}]\right)\eta_{\hat{y}}}{\sum_{\hat{y}=1}^{C}\exp\left(\eta_{\hat{y}}^T\,\mathbb{E}_q[\bar{z}]\right)}\right) \tag{6}$$

The last term of equation 6 can be written using $s = \mathrm{softmax}(\mathbb{E}_q[\bar{z}], \eta)$ as $\eta_y - \sum_{\hat{y}=1}^{C} s_{\hat{y}}\eta_{\hat{y}}$. Experiments show that s changes very slowly with respect to ϕ_n and therefore we derive the subsequent closed form update rule.

$$\phi_n \propto \beta_n \exp\left(\Psi(\gamma) + \frac{1}{N}\left(\eta_y - \sum_{\hat{y}=1}^{C} s_{\hat{y}}\eta_{\hat{y}}\right)\right) \tag{7}$$

Using the update rule from equation 7 we managed to alleviate the computational problems indicated in 2.1. We thus attain comparative computational complexity with the unsupervised LDA, while preserving the supervised part. However, the supervised part in this update rule is less dominant than the unsupervised one, due to the multiplication with the $\frac{1}{N}$ factor.

In order to address this problem we introduce the update rule shown in equation 8, where $\mathcal{C}$ is a free to change hyperparameter, which can influence the effect of the supervised part on the topics inference.

$$\phi_n \propto \beta_n \exp\left(\Psi(\gamma) + \frac{\mathcal{C}}{\max(\eta)}\left(\eta_y - \sum_{\hat{y}=1}^{C} s_{\hat{y}}\eta_{\hat{y}}\right)\right) \tag{8}$$

Intuitively, this update rule is justifiable if we consider that what actually matters is the relative proportions of the two terms and not the magnitude of their values, since ϕ_n is normalized. The update rule for γ remains the same as in the unsupervised LDA.

In the maximization step, we need to maximize $\mathcal{L}$ with respect to the model parameters β and η. The maximization with respect to the topics β remains the same as for the unsupervised LDA. To compute the classification parameters η we need to maximize $\mathcal{L}_\eta$.

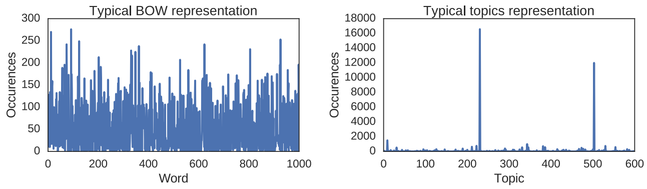

Figure 2: Typical Bag of Words and topics representations for a video

$$\mathcal{L}_\eta = \sum_{d=1}^{D}\eta_{y_d}^T\,\mathbb{E}_q[\bar{z_d}] - \sum_{d=1}^{D}\log\sum_{\hat{y}=1}^{C}\exp(\eta_{\hat{y}}^T\,\mathbb{E}_q[\bar{z_d}])\left(1+\frac{1}{2}\eta_{\hat{y}}^T\,\mathbb{V}_q[\bar{z_d}]\eta_{\hat{y}}\right) \tag{9}$$

We have already mentioned that the influence of $\mathbb{V}_q[\bar{z}]$ is insignificant. In addition, in order to avoid computing and keeping the variance matrix in memory for every document, we adopt the first order approximation which amounts to multinomial logistic regression with respect to $\mathbb{E}_q[\bar{z}]$.

3. EXPERIMENTS

In this section, fsLDA is evaluated in two action recognition datasets of realistic videos. We assess its performance regarding two metrics, namely discriminativeness of the topic mixture representations and qualitative assessment of the semanticness of the inferred topics.

3.1 Experimental setup

Two datasets are used in our experiments, the UCF11-Youtube Action Dataset [5] and the UCF101-Action Recognition Dataset [12]. *UCF11* is composed of 11 action classes with 1600 videos, the majority of which contain heavy camera motion. *UCF101* is one of the state-of-the-art datasets for action recognition. It consists of 13320 videos, belonging to 101 categories.

We use both visual and motion features, in the conducted experiments, to establish that the proposed method is effective regardless of the nature of the local features. In order to represent motion information, Improved Dense Trajectories (IDT) [14] are extracted from each video for both datasets. In case of UCF11, Dense SIFT [6] are computed to encapsulate visual information. Regarding encoding visual information from UCF101, we decided to use the last two convolutional layers from [11] as local features in $\mathbb{R}^{512}$. The extracted local features are subsequently encoded using Bag of Words (BoW) representation with 1000 and 4000 codewords for UCF11 and UCF101 respectively.

In subsequent experiments, we measure the classification performance of a linear SVM by computing the mean accuracy score in three random splits. C for the SVM is chosen via cross-validation while the $\mathcal{C}$ hyperparameter of our method is selected by hand. We compare our method with both supervised and unsupervised LDA as well as with BoW.

3.2 Qualitative topic evaluation

It simply suffices to observe the topic versus word representation in Figure 2 to notice that topics are by far more sparse than words. Intuitively, this can be attributed to the topics encapsulating much more information than a sin-

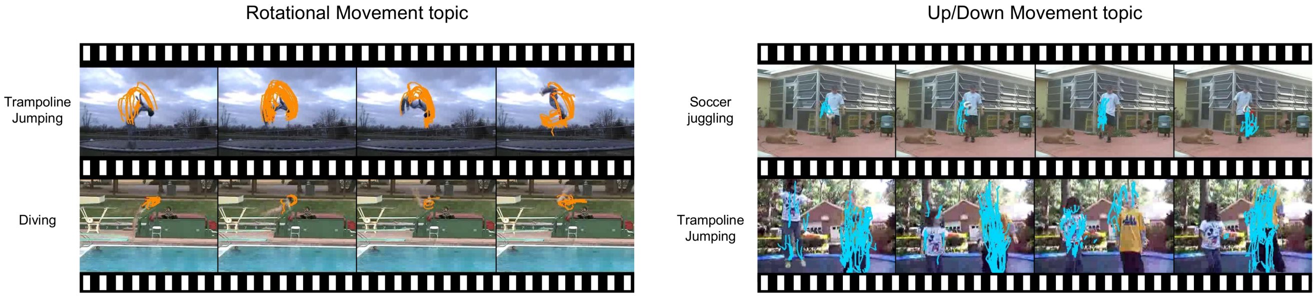

Figure 3: Visualization of rotational movement topic (orange) and vertical movement topic (light blue). The red bullet indicates the beginning of the trajectory.

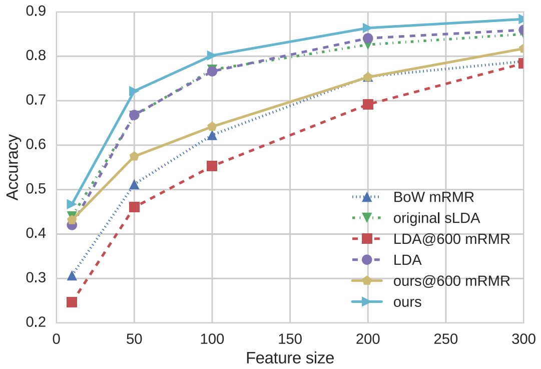

Figure 4: Classification accuracy using few dimensions to represent videos

gle word. In Figure 3, we draw the IDT trajectories that correspond to a codeword that belongs in the ten most common words of two different topics. It is noticeable that the inferred topics capture the specifics of different complex motions, such as rotational and vertical movements. These motions are discovered in videos from different classes, such as trampoline jumping and diving for rotational movement and soccer juggling and trampoline jumping for vertical movement. We have observed that many topics discovered with unsupervised LDA refer to background information, which is the dominant structure in the video. In contrast, the topics inferred using our method encapsulate high level information, which corresponds to micro-events.

In order to evaluate the class-relevant information encoded in each topic, we reduce feature dimensionality using either minimum Redundancy Maximum Relevance Feature Selection (mRMR) [7] or by simply training our method with a smaller number of topics. Figure 4 depicts the classification performance according to the aforementioned procedure on UCF11 using a single representative descriptor (idt-hog). Each of the topics discovered by fsLDA contains more class-relevant information compared to the ones inferred by both LDA and sLDA as well as to the codewords found by KMeans. This is established since the proposed method outperforms BoW using mRMR feature selection. The same results are observed when selecting features with other feature selection methods.

3.3 Evaluation of discriminativeness of topics

In this section, we evaluate the discriminativeness of the inferred topic distributions by measuring the classification performance on a variety of descriptors using the whole topic distribution or the concatenation of the topic distributions (denoted as combined) as a feature.

Figure 5 depicts the mean accuracy for every descriptor in UCF11. Our method outperforms LDA in every descriptor and sLDA in all but one while being 50 to 100 times faster. Table 1 presents the three-fold accuracy scores in the case of UCF101 for some indicative descriptors. Even in this more demanding large-scale dataset, fsLDA performs better in terms of classification accuracy especially for the conv5_2 features.

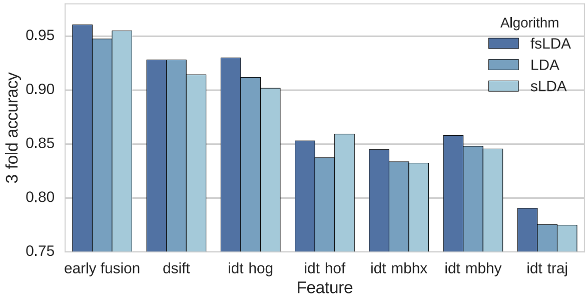

Figure 5: Comparison of fsLDA with unsupervised LDA and sLDA using 600 topics in UCF11

	LDA @ 1200	fsLDA @ 1200
conv5_2	56.03%	**62.37%**
idt-hof	52.72%	**56.07%**
combined	67.50%	**69.87%**

Table 1: Comparison of fsLDA with unsupervised LDA in UCF101 (the complexity of sLDA for UCF101 is prohibitive)

4. CONCLUSIONS

We have developed a new method to infer topics in a supervised manner which, in contrast to sLDA, is tractable on large-scale video datasets such as UCF101. Furthermore, we have shown that the proposed method outperforms unsupervised LDA and discovers topics which encapsulate high level information corresponding to micro-events, while containing more class-relevant information than words.

Future work includes the study of topics as rich descriptions of video attributes for video captioning and retrieval applications.

5. REFERENCES

[1] D. M. Blei and J. D. Mcauliffe. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128, 2008.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] M. Dixit, N. Rasiwasia, and N. Vasconcelos. Class-specific simplex-latent dirichlet allocation for image classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2672–2679, 2013.

[4] R. R. Iyer, S. Parekh, V. Mohandoss, A. Ramsurat, B. Raj, and R. Singh. Content-based video indexing and retrieval using corr-lda. *arXiv preprint arXiv:1602.08581*, 2016.

[5] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos in the wild. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1996–2003. IEEE, 2009.

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[7] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

[8] D. Putthividhya, H. T. Attias, and S. S. Nagarajan. Supervised topic model for automatic image annotation. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 1894–1897. IEEE, 2010.

[9] S. Qian, T. Zhang, C. Xu, and M. S. Hossain. Social event classification via boosted multimodal supervised latent dirichlet allocation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(2):27, 2015.

[10] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[11] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] 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.

[13] C. Wang, D. Blei, and F.-F. Li. Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1903–1910. IEEE, 2009.

[14] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3551–3558, 2013.

[15] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, pages 461–470. ACM, 2015.

[16] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4694–4702, 2015.

[17] J. Zhu, A. Ahmed, and E. P. Xing. Medlda: maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th annual international conference on machine learning*, pages 1257–1264. ACM, 2009.