A Combination of Generative and Discriminative Models for Fast Unsupervised Activity Recognition from Traffic Scene Videos

Mahesh Venkata Krishna and Joachim Denzler
Computer Vision Group, Friedrich Schiller University Jena
Ernst-Abbe-Platz 2, 07743 Jena, Germany
{mahesh.vk,joachim.denzler}@uni-jena.de

Abstract

Recent approaches in traffic and crowd scene analysis make extensive use of non-parametric hierarchical Bayesian models for intelligent clustering of features into activities. Although this has yielded impressive results, it requires the use of time consuming Bayesian inference during both training and classification. Therefore, we seek to limit Bayesian inference to the training stage, where unsupervised clustering is performed to extract semantically meaningful *activities from the scene. In the testing stage, we use discriminative classifiers, taking advantage of their relative simplicity and fast inference. Experiments on publicly available data-sets show that our approach is comparable in classification accuracy to state-of-the-art methods and provides a significant speed-up in the testing phase.*

1. Introduction

Inferring *spatio-temporal dependencies* is a major challenge in machine vision. As providing training labels is often difficult or impractical in these situations, it becomes imperative to use unsupervised learning approaches. Significant progress has been made over the years on the problem of unsupervised scene analysis and event detection (some early examples can be seen in [16, 8]). The problem is not only theoretically important, but also has many practical applications, such as traffic scene analysis, crowd behavior analysis *etc*. [19, 9, 6].

We concentrate here on the application of unsupervised traffic scene analysis (cf. Fig. 1). In these videos, activities happening in the scene are characterized by the relationships (both spatial and temporal) between the movements of the contained objects. The main issues here are the complexity of the scenes and the high variability of objects in terms of appearance. There are often more than a hundred objects in the scene and it is nearly impossible to use simple object tracking approaches to analyze these situations due to heavy occlusions. As can be seen from Fig. 1, simple clustering and classification schemes will not be able to handle the activities at the junction.

Figure 1. Examples of typical traffic scenes. Some of the activities have been marked with arrows. Activities are characterized by the spatio-temporal relationships between the motions of the actors.

The dynamic complexity of the scene motivates us to use low-level features such as optical flow and then use non-parametric methods to cluster them into meaningful activities. A frequently adopted approach to achieve this is to use *probabilistic topic models* [19, 9, 7]. These models can extract meaningful latent structures from the input data. However, a major shortcoming is the time complexity of these models, especially when new large-scale video streams need to be analyzed. Therefore, the main contribution of this paper is a novel way of combining *generative* (in this case the *Hierarchical Dirichlet Processes (HDP)* models) and *discriminative* models in a step-wise manner to reduce computational complexity and running time during testing. This allows for using models learned from HDP for traffic scene analysis. Furthermore, previous studies provided only a limited quantitative evaluation of these strategies mostly tested on data-sets with only very basic activi-

ties (*e.g.*, horizontal and vertical motion used in [10, 11]). In contrast, our analysis is based on three publicly available data-sets with one containing seven different activities.

1.1. Previous Work

Li *et al*. [10] and Hospedales *et al*. [6] demonstrated the possibility of using topic models from natural language processing domain for video analysis. Further works by Kuettel *et al*. [9] and Wang *et al*. [19] established the effectiveness of HDPs for video scene analysis with an emphasis on crowd and traffic scenes. There are other approaches in the same line, *e.g*. [7, 11], which use various adaptations of the topic models to the problem at hand. The main step in these approaches is the *Bayesian inference* of latent parameters, where the observed words in documents are clustered into topics based on co-occurrence. A major stumbling block here is that the Bayesian inference step is computationally expensive, in spite of attempts such as that by Yao *et al*. [20] to reduce this overhead. The complexity mainly arises from the fact that each sampling iteration requires a large number of operations to be performed, including random number generation.

Another interesting method tackling the problem is the one by Ricci *et al*. [14]. The authors extract activity histograms from short motion trajectories and then use the *Earth Movers Distance* (EMD) as the objective function to solve the resulting matrix factorization problem for clustering. However, when one needs to consider a large dimensionality of histograms or a large number of words, the optimization problem can be cumbersome and time consuming. In such cases, topic models are a better alternative and our approach, as we will see, handles such situations well and provides comparable performance.

In the works by Nater *et al*. [12, 13], simple tracker hierarchies are arranged in a tree-structure to analyze human behavior. However, they only consider temporal interdependencies and hence can not separate two atomic activities happening at the same time. Our approach, being based on the HDP model, inherently clusters based on spatio-temporal co-occurrences and is therefore a more representative model of the scene. Fritz *et al*. [4] also use generative models to initially cluster features into classes, but their approach requires additional knowledge about the system, like the number of target classes.

We use topic models to extract activities from the observed motion vectors and use these for training a discriminative classifier. This gives us the full capabilities of HDPs in modeling complex scenes with the speed of discriminative classification approaches during testing time. We will also demonstrate that our approach performs at least on par with the state-of-the-art with respect to recognition accuracy.

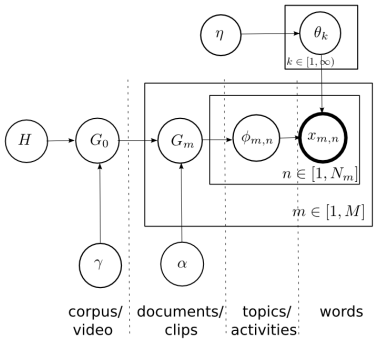


Figure 2. A complete HDP model formulation (from a stick-breaking construction perspective)

2. Hierarchical Dirichlet Processes

Modeling principles : The use of HDPs for modeling machine learning data followed the seminal work by Teh *et al*. [15]. These models can be viewed as a generalization of Latent Dirichlet Analysis (LDA) [2], but unlike LDA, in HDP models, the number of latent topics is *inferred* from the data and a small number of hyper-parameters. Figure 2 shows the basic HDP model. Suppose we are given an input data *corpus*, which is split into M groups (commonly referred to as *documents*) and each document contains N_m data points $x_{m,n}$(called *words*). It is then the goal of the HDP model to cluster these words into meaningful latent structures (called *topics*). It is worth noting that the order/position of the appearance of the words within a document is not considered inherently by this approach.

For video processing, we follow [9, 19], *i.e*., a corpus corresponds to the whole video, documents correspond to clips, topics correspond to atomic activities and words correspond to the 4-tuple of optical flow and position coordinates. We use these pairs of terms interchangeably in this work.

In an HDP, a Dirichlet Process(DP) generates the global, video-wide list of activities. Then, for each clip, we have a DP generating the list of activities in that particular clip. These clip-specific DPs are seen to be drawn from the global list. Formally, we write the generative HDP formulation as follows.

$$\begin{aligned} G_0 \,|\, \gamma, H &\sim DP(\gamma, H) \\ G_m \,|\, \alpha, G_0 &\sim DP(\alpha, G_0) \quad \text{for } m \in [1, M] \end{aligned} \tag{1}$$

The hyper-parameters γ and α are called the concentration parameters and the distribution H is called the base distribution (which, in our case, is Dirichlet distributed with a parameter D_0). Next, the observed words $x_{m,n}$ are seen as being sampled from the mixture priors $\phi_{m,n}$, which in turn are seen as being drawn from a Dirichlet Process G_m. The possible values for mixture components are drawn from

another process, θ_k. Thus, the remaining part of the formulation of this construction can be written as,

$$\begin{aligned} \theta_k &\sim P(\eta) \quad \text{for } k \in [1, \infty) \\ \phi_{m,n} \,|\, \alpha, G_m &\sim G_m \quad \text{for } m \in [1, M], n \in [1, N_m] \\ x_{m,n} \,|\, \phi_{m,n}, \theta_k &\sim F(\theta_{\phi_{m,n}}) \end{aligned} \tag{2}$$

Here, M is the number of clips, N_m is the number of words in clip m, $P(\cdot)$ is the prior distribution over topics and $F(\cdot)$ is the prior word distribution given the topic.

Bayesian Inference : In our task, we have to tackle the problem of *Bayesian Inference*, *i.e.*, given $x_{m,n}$, how to calculate $\phi_{m,n}$? In general, this is a hard problem and closed form expressions do not exist for the target distributions. Hence, we need to use approximate methods, like the *Markov Chain Monte Carlo* (MCMC) methods ([1, 5]), especially Gibbs sampling. Using the well-known Chinese Restaurant Franchise-based formulation, we obtain the following expression for sampling the conditional distribution for Gibbs sampler iteration ([15]):

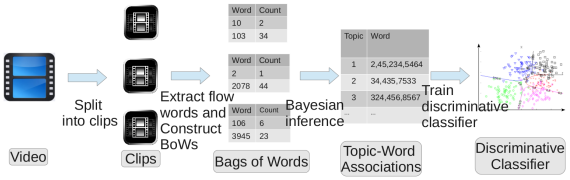
$$\begin{aligned} p(\phi_{m,n} = k \,|\, x, \alpha, \gamma, \eta, \theta, H) \propto \\ (n_{m,k}^{\neg m,n} + \alpha\theta_k) \cdot \frac{n_{k,t}^{\neg m,n} + \eta}{n_k^{\neg m,n} + V \cdot \eta} \end{aligned} \tag{3}$$

where $n_{m,k}, n_{k,t}$ and n_k represent count statistics of the word-topic associations, topic-document associations and the topic-wise word counts, respectively. The superscript $\neg m, n$ indicates that the present word, $x_{m,n}$, is to be excluded from these statistics. V is the size of the dictionary.

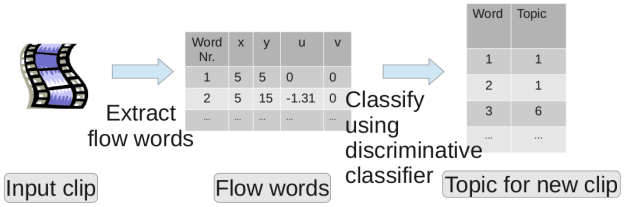
The first term in equation (3) implies that, the probability that the current word will be associated with a topic, is proportional to the number of words already assigned to that topic. The second term (which is the probability of starting a new topic) shows that the hyper parameters α, γ and especially η can be used to control the number of topics inferred. We also perform hyper-parameter sampling to make our framework completely data-driven, without any supervision.

3. Fast Activity Classification

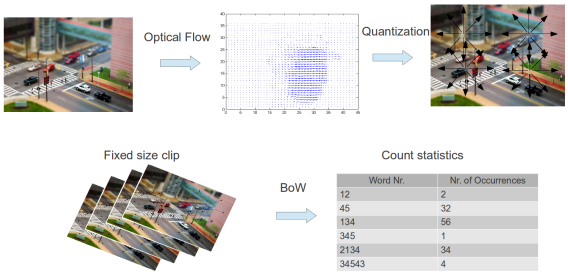
Our approach is a step-wise combination of generative and discriminative approaches. In the first step, we take a training video (split into M clips), from which activities θ_k are extracted. They are then assigned to flow-words $x_{m,n}$ by mixture variables $\phi_{m,n}$. In the second step, we use the set of pairs $(x_{m,n}, \phi_{m,n})$ obtained from the generative model as the training data for a discriminative classifier. Once such a classifier is trained, in the third step, new unseen incoming video data is analyzed using the discriminative classifier trained previously. We now describe the process in more detail. Figure 3 summarizes our approach.



(a) Training



(b) Testing



(c) Process of extracting a bag-of-flow words representation for each clip

Figure 3. Summary of our step-wise approach.

From the input video, we extract the optical flow for each consecutive pair of frames. For this, we use the publicly available implementation of the TV-L[1] algorithm [21]. Next, the optical flow vectors are thresholded to remove noise and are quantized into 8 directions. A dictionary is then built with all possible flow words (flow words are four-tuples, with the x-y co-ordinates and associated flow values). Then, we divide the video into clips of equal size (of a few seconds) and form a bag-of-words representation for each clip. Figure 3(c) shows this process.

Next, we perform Bayesian inference for HDP. The hyper-parameters are set to control the total number of topics extracted. For each word in the input data, we get one topic associated with it. Though in theory we can have an infinite number of topics, we consider only the activities which explain at least 5% of the scene. The others are often insignificant topics, containing too few flow words to explain any meaningful activity in the scene and arise most often due to noisy flow features.

The next step is the training of a discriminative classifier. For this purpose, we use a one-versus-all C-SVM classifier (using [3]) with a Gaussian Radial Basis Function(RBF) kernel (similar to [4]). Since we have multiple classes, we

train K (number of topics) binary C-SVM classifiers and vote for the final classification. Flow words of each class form the positive samples for it and the rest form the negative samples. Finally, the incoming video is grouped into clips and optical flow words are extracted, which are then classified by the SVM classifier.

The activities happening in the scene are determined by voting. Not all activities are present in all image regions. Thus, to boost the performance of the classifier, we split each frame into 4 quadrants before training, which reduces classification problem size by restricting the classification problem to the activities within the respective quadrants.

Abnormality Detection: For abnormality detection, we take the classification results of the above SVM classifiers and simply declare non-conformants as abnormalities. For example, if the classification result declares that 2 activities are happening in a clip, then, any flow vector that does not belong to these activities is an abnormality (we ignore single outliers as noise and consider only groups of a minimum of 4 flow words). This method, while simple, yields satisfactory results, as will be shown later. As can be expected, the gain in computation speed and reduction of algorithm complexity is quite considerable, when contrasted with another inference step required by purely generative methods. There, one has to perform another sampling step to determine the probability that the new document was generated by the learned model and then mark as abnormality the clips with very low probability.

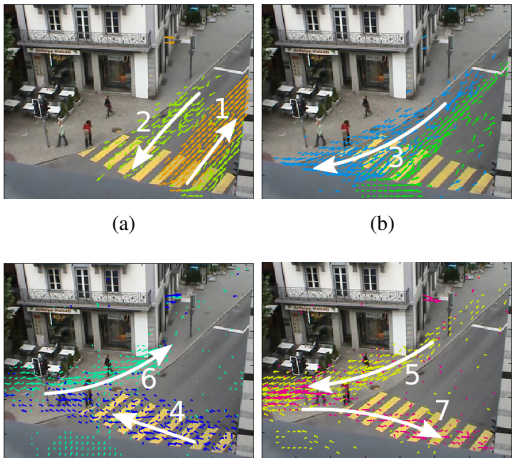
4. Experiments and Results

For validating our approach, we have used three different publicly available data-sets. The Bayesian inference during training was realized with a Gibbs' sampler. To speed up computations and to reduce the number of iterations required, we used the techniques proposed in [20]. We also used the split-and-merge procedure described in [18]. As mentioned before, hyper-parameter were sampled to make them driven by data, but as initial values, we set $\alpha = 1$, $\gamma = 1$ and $\eta = 0.5$ in all cases.

4.1. Complex Activity Extraction and Classification

Data-set : In the relevant previous works like [11, 14], one major drawback in assessing true performance is that they generally use data-sets with ground truth labels such as "horizontal" or "vertical" traffic flow. This leads to a relatively simple recognition problem that does not demonstrate the true capabilities of the respective methods in extracting complex meaningful activities from the scene in an unsupervised manner.

To overcome this problem and to demonstrate the capability of our fast HDP techniques, we use the *Traffic* dataset of [17]. This video contains over 44 minutes of recording of road traffic with a resolution of 360×288 pixels per frame.



(a) (b) (c) (d)

Figure 4. Activities extracted by the HDP model for the *Traffic* sequence. The arrows in white show the ground truth activities. (Split into four images for better viewing)

The video was split into clips of 10 seconds each, to obtain 265 clips. We divided them into 100 training and 165 testing clips. A human annotator marked the salient and *semantically meaningful* activities in the scene, at the scene level (not at the flow-word level)[1]. There were 7 activities in the ground truth, 3 vehicular and 4 pedestrian. Figure 4 shows an example frame with ground truth activities marked with the thick arrows.

Results : To qualitatively analyze the output of our method, we plot the activities extracted by the HDP model for the *Traffic* sequence in Figure 4. As can be seen, these extracted activities closely match the human ground truth markings, and it shows that our method can extract complex and semantically meaningful activities that simple clustering/classification schemes can not extract and classify.

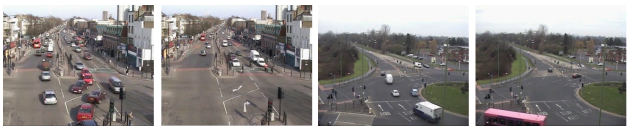
We compare our results with the purely generative approach of [9], using the code made public by the authors. Table 1 shows the results. Our approach performs better in terms of accuracy, and, more importantly, in terms of run time during testing.

To compare our method with a baseline cluster/classify scheme (without supervision) and to demonstrate where such simpler schemes fail in the present application, we used Gaussian Mixture Model (GMM) clustering with 7 cluster points. Classification was performed as before using the SVM classifier learned from the extracted clusters. The accuracy for this method was only 45.43% and it failed where activities were mainly consisting of diverse flow directions or overlapped spatially.

[1]Ground-truth link: https://cms.rz.uni-jena.de/bin/page/index.php?id=710&suffix=zip&nonactive=1&lang=de&site=dbvmedia

Table 1. Comparison of our approach with that of [9] on the *Traffic* dataset.

Data-set	Accuracy	Training Time	Testing Time
DDP-HMM [9]	87.88%	26 min	31 min
Our Approach	**92.12%**	**12 min**	**4 secs**



(a) *Junction* (b) *Roundabout*

(c) *MITTraffic*

Figure 5. Example frames of the *Junction*, *Roundabout* and *MIT-Traffic* sequences.

4.2. Comparison with State-of-the-Art

To demonstrate that our two-step method performs on par with state-of-the-art even while providing significant gains in terms of reduced complexity and run time in the testing stage, we use publicly available data-sets with marked ground truth. In order to make the comparison meaningful, we used parameters consistent with previous works in extracting and quantizing optical flow, with 8 flow directions [9]. We compare our method with [14, 11, 9]. For [9], we used the code made available by the authors. For [14, 11, 9], a direct comparison with the their results is possible on the following data-sets.

Junction and Roundabout data-sets : The *Junction* and *Roundabout* data-sets ([11, 10]) are made up by 33600 and 61500 frames, respectively, each of size 360 $\times$ 288 pixels. The videos are divided into clips of 12 seconds each. Example frames can be seen in Fig. 5(a) and 5(b). There are 8 abnormal activities in the *Junction* video and 6 in the *Roundabout* video.

In these data-sets, the density of vehicles is comparatively high and motion is, therefore, quite complex. The ground-truth data is supplied with the data-sets (with two simple labels, *i.e.*, horizontal and vertical traffic flow).

MITTraffic data-set : The *MITTraffic* dataset ([19]) contains 20 video recordings of a traffic junction. The first video has 8295 frames and the subsequent videos have 6920 frames each. The frame size is 720 $\times$ 480. We split the videos into clips of 8 seconds each. Figure 5(c) shows example frames for this sequence. As the ground-truth data was unavailable for this data-set, they were marked in a manner consistent with the previously mentioned data-set. We divided the traffic flow into horizontal and vertical directions and used that as the ground-truth for the clips.

Figure 6. Six most probable activities extracted for the *Junction* dataset.

Table 2. Comparison of classification accuracy

Data-set	EMD-L_1 [14]	Cas-pLSA [11]	DDP-HMM [9]	Our Approach
Junction	**92.31%**	89.74%	87.18%	**92.31%**
Roundabout	86.4%	76.2%	83.05%	**88.13%**
MITTraffic	NA	NA	84.21%	**89.47%**

Results : Figure 6 shows the actions extracted for the *Junction* sequence. Even though our method extracts many complex activities from the scene (as shown in Fig. 6), due to limitations in data-set ground truth (and to enable effective comparison with [14, 11]), quantitative evaluation is performed by mapping the perceived activities to two-classes in the case of *Junction, Roundabout* and *MITTraffic* videos depending on the mean flow direction. Table 2 summarizes the results.

The results show that our algorithm achieves same levels of performance compared to other relevant works. The real gain, however, is in run time. For the *MITTraffic* sequence, the completely generative approach of [9] took 38 minutes to train and 41 minutes to test, whereas our approach only required 12 minutes and 5 seconds respectively.

In the baseline GMM-based approach of Sec. 4.1, with the number of clusters set to 2 (for horizontal and vertical), we achieved a classification accuracy of 97.36% for *Junction* data-set, which demonstrates that the simplistic ground truth provided with the *Junction* and *Roundabout* data-sets is not sufficient to quantitatively analyze the capabilities of activity extraction and classification systems.

4.3. Abnormality Detection

We performed abnormality detection for the *Junction* and *Roundabout* data-sets using their ground truth. For the *Junction* video, we achieved True Positive Rate (TPR) of 75% at a False Positive Rate (FPR) of 12.9% and for the

Figure 7. Detected abnormality in clip 4 of the *Junction* data-set, where a fire-engine interrupts traffic flow. (Some flow vectors have been removed to improve visibility.)

Roundabout sequence, we achieved TPR of 83.33% at a FPR of 22.64%. These results compare favorably with those reported in [11] (TPR of 75% at FPR of 12.9% for *Junction* and TPR of 83.33% at FPR of 33.9% for *Junction* video). Figure 7 shows an abnormality detected in a clip from the *Junction* sequence, caused by a fire engine interrupting the normal traffic flow.

5. Conclusions

We have presented a novel combination of generative and discriminative models for the purpose of unsupervised traffic scene analysis and fast analysis of large-scale video data. Results demonstrated that our method is capable of perceiving meaningful activities from complex traffic video data. We also showed that while this method achieves state-of-the-art performance in both classification accuracy and abnormality detection, it provides significant speed up in computation time by approximating Bayesian inference during testing with a margin-based classifier.

We have considered the application of traffic scenes in this work. However, our approach is not limited to this application and other feature representations could be used in this setup. We hope that the speed up provided by our method will widen the scope of HDP models to other video scenes and application fields. Furthermore, currently, the HDP models can not handle temporal ordering information. Developing methods that can incorporate temporal information into HDPs might be very useful in analyzing repetitive activities in traffic scenes.

Acknowledgements The first author is supported by the Carl Zeiss AG through the "Pro-Excellence" scholarship of the State of Thuringia, Germany. The authors are thankful to Erik Rodner, Christian Wojek and Stefan Saur for useful discussions and suggestions.

References

[1] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to mcmc for machine learning. *Machine Learning*, 50(1):5–43, January 2003. 3

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

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 3

[4] M. Fritz and B. Schiele. Decomposition, discovery and detection of visual categories using topic models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, 2008. 2, 3

[5] W. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics*. Chapman and Hall/CRC, 1995. 3

[6] T. Hospedales, S. Gong, and T. Xiang. A markov clustering topic model for mining behaviour in video. In *IEEE 12th International Conference on Computer Vision*, pages 1165–1172, 2009. 1, 2

[7] T. Hospedales, J. Li, S. Gong, and T. Xiang. Identifying rare and subtle behaviors: A weakly supervised joint topic model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2451 –2464, dec 2011. 1, 2

[8] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 34(3):334–352, 2004. 1

[9] D. Kuettel, M. D. Breitenstein, L. V. Gool, and V. Ferrari. What is Going on? Discovering SpatioTemporal Dependencies in Dynamic Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2010. 1, 2, 4, 5

[10] J. Li, S. Gong, and T. Xiang. Global behaviour inference using probabilistic latent semantic analysis. In *British Machine Vision Conference*, pages 193–202, 2008. 2, 5

[11] J. Li, S. Gong, and T. Xiang. Learning behavioural context. *International Journal of Computer Vision*, 97(3), may 2012. 2, 4, 5, 6

[12] F. Nater, H. Grabner, , and L. V. Gool. Exploiting simple hierarchies for unsupervised human behavior analysis. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2010. 2

[13] F. Nater, H. Grabner, and L. V. Gool. Temporal relations in videos for unsupervised activity analysis. In *British Machine Vision Conference*, 2011. 2

[14] E. Ricci, G. Zen, N. Sebe, and S. Messelodi. A prototype learning framework using emd: Application to complex scenes analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):513–526, 2013. 2, 4, 5

[15] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, pages 1566–1581, 2006. 2, 3

[16] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473 –1488, nov. 2008. 1

[17] J. Varadarajan, R. Emonet, and J.-M. Odobez. Probabilistic latent sequential motifs: Discovering temporal activity patterns in video scenes. In *BMVC*, pages 1–11, 2010. 4

[18] C. Wang and D. M. Blei. A split-merge mcmc algorithm for the hierarchical dirichlet process. *ArXiv e-prints, arXiv:1201.1657 [stat.ML]*, Jan 2012. 4

[19] X. Wang, X. Ma, and W. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(3):539 –555, march 2009. 1, 2, 5

[20] L. Yao, D. Mimno, and A. McCallum. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 937–946, 2009. 2, 4

[21] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l1 optical flow. In *Pattern Recognition (Proc. DAGM)*, pages 214–223, 2007. 3