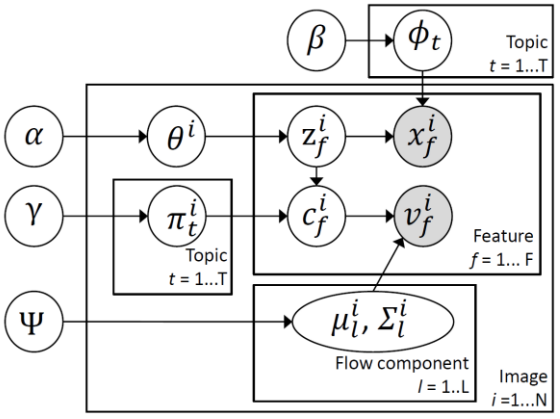
Learning image models using 'flobject analysis'

Inmar E. Givoni*, Patrick Li*, Brendan J. Frey
University of Toronto
{inmar,pli,frey}@psi.toronto.edu
http://www.psi.toronto.edu/

A promising direction of vision research is to develop unsupervised learning algorithms that can extract concise representations of images. These representations can then be used as inputs for supervised learning tasks, such as object classification, scene recognition and image segmentation. The most successful vision systems for interpreting static images that we know of are biological ones. Interestingly, these systems have access to motion and/or stereo information that are likely instrumental for learning good image representations. We therefore ask the question: *Can motion and/or stereo disparity information be used to train better methods for extracting representations from static images?* To answer this question, we explore an analogous framework for unsupervised learning. The output of our method is a model that can generate a vector representation or descriptor from any static image. However, the model is trained using pairs of consecutive video frames, which are used to find representations that are consistent with optical flow-derived objects, or 'flobjects'.

FLDA model: To demonstrate the flobject analysis framework, we extend the latent Dirichlet allocation bag-of-words model [1] to account for real-valued word-specific flow vectors and image-specific probabilistic associations between flow clusters and topics. This model, denoted by FLDA, is trained using the optical flow, but is applicable in its absence. FLDA can account for appearance and flow such that visual features associated with similar flow are more likely to belong in the same topic and visual features associated with dissimilar flow are more likely to belong in different topics. In addition, FLDA can account for multiple moving objects in one image including different objects that exhibit similar velocity and a single object that exhibits different velocities (as in the case of articulation).



x – appearance feature
v – flow feature (observed only during FLDA learning)
c – flow component
z – topic
θ – multinomial distribution over topics
π – multinomial distribution over flow components
ϕ – multinomial distribution for topic over codebook
μ, – normal distribution for flow motion parameters
γ – Dirichlet prior parameters for flow component distribution
α, – Dirichlet prior parameters for image and topic distribution
Ψ – Normal-Inverse Wishart prior for flow parameters.

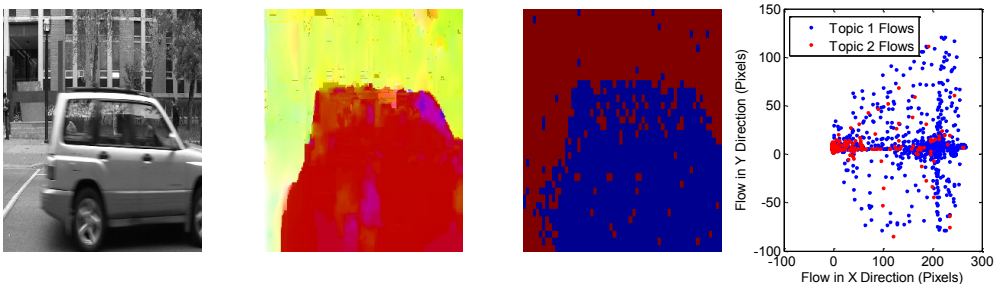
The above figure describes the graphical model for FLDA. The topics are learned using collapsed Gibbs sampling. Then, given a static image and the learned topics, each visual feature in the image is assigned a topic, and these assignments are summarized in concatenated topic-

* These authors contributed equally.

Topic: vision, learning algorithms , **Preference: oral** , **Presenting Author: Givoni or Frey**

specific visual-word histograms that comprise the FLDA based descriptor. This descriptor can then be used for various tasks, in a supervised learning setting.

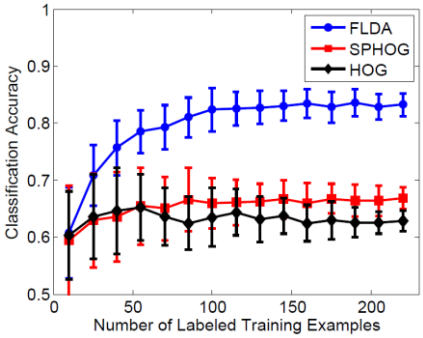
Dataset for FLDA: We created two new challenging datasets of image pairs extracted from videos, which we use for training the unsupervised model. One dataset consists of rigid objects (cars) and the other consists of articulated objects (pedestrians). The set of figures below show, from left to right, the first image in an image pair, the flow calculated based on the image pair, the FLDA topic assignments for each extracted visual feature, and a scatter plot of the flow colored by the FLDA topic assignment.



Experiments: We use the static image FLDA descriptors for a simple classification task: to determine whether an image contains a car or not using nearest neighbors. We compared classification performance of various image descriptors (HOG, spatial pyramid HOG (SPHOG) [2], Gist, standard LDA, and our method, FLDA) using different histogram normalization schemes. In the image set below the left column shows a test image, the middle column shows the FLDA based descriptor nearest neighbor, and the right column shows the SPHOG nearest neighbor. The table shown below demonstrates that FLDA achieves higher classification rates compared to the other descriptors for the car dataset. We obtain similar results for intersection kernel NN classifier, and for the pedestrian dataset.

NN	None	L1	L2
HOG	**65%**	60%	54%
SPHOG	**65%**	64%	54%
Gist	69%	69%	**70%**
LDA	62%	**64%**	59%
FLDA	61%	**82%**	73%

The plot shown to the right describes how the performance of HOG, SPHOG and FLDA based descriptors is affected by the number of labeled training examples used. When the number of training samples is small all methods perform similarly, but as more training examples are added, FLDA can better utilize the additional information.



[1] DM Blei,AY Ng, M. Jordan. *Latent Dirichlet Allocation*, 2003
[2] A Oliva, A Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope, 2001.