Object Categorization Using Spatial-Temporal Features

Chrisil Arackaparambil Ashok Chandrashekhar

1 Introduction

Objects from different semantic categories often have different motion signatures. Humans, animals, birds, cars, airplanes etc, have very distinct patterns of motion. Therefore, along with spatially salient features, the temporal patterns of movement can be used as features in order to determine the category of an unlabeled object.

In our project, we explore the suitability of spatio-temporal features in order to perform unsupervised classification of a set of objects belonging to different semantic categories using videos. That is, we use motion along with spatial properties to determine object categories. Towards this goal, we use the feature extraction algorithm of Dollár et al [1] to detect and isolate the salient spatio-temporal features in videos and use these features with a Bag-of-Features model to perform classification.

2 Feature Extraction Algorithm

The first step in our approach is that of feature extraction: we need to get a suitable representation of features corresponding to different categories, and be able to extract them from the videos in our dataset. We use the algorithm of Dollár et al [1] that produces a set of feature vectors when given an input video. These feature vectors identify the salient complex features that we need to capture.

The algorithm is a linear separable filter that computes the following response function:

$$R = (I \star g \star h_{\text{ev}})^2 + (I \star g \star h_{\text{od}})^2$$

where, $g(x, y)$ is the 2D Gaussian smoothing kernel, and $h_{\text{ev}}(t)$ and $h_{\text{od}}(t)$ are a quadrature pair of 1D Gabor filters. This function is designed to produce a strong response for complex spatio-temporal motions. An assumption made with this filter is that the camera is stationary, and we will see how our results are affected when this assumption does not hold.

The algorithm determines the local maxima of the response function in the space-time domain, and these points are identified as interest points. Figures 1 and 2 show the response induced due to an input sequence from our data set.

A space-time region around an interest point is called a *cuboid*. Thus, we end up with a set of cuboids for the given input video. To reduce the dimension of the features that are output, first the algorithm computes a brightness gradient for each cuboid, and applies PCA on the vectorized versions of the gradients. These resulting vectors are taken to be the feature vectors of the input video, for use in classification. Figure 3 shows the interest points in the input sequence considered earlier, and Figure 4 shows some of the resulting cuboids.

Figure 1: A sequence from one of the videos in our dataset

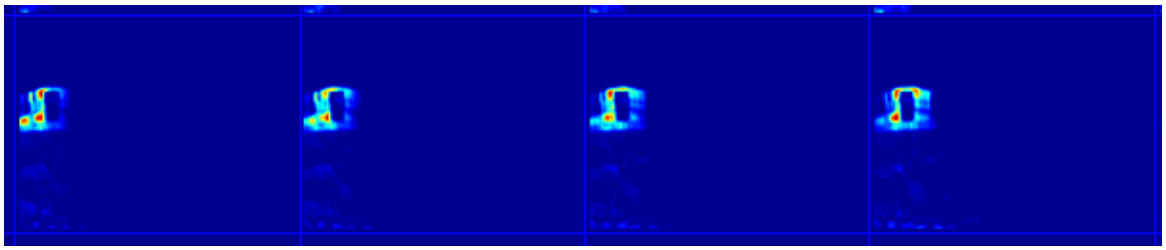

Figure 2: The response induced due to the input sequence

3 Classification

3.1 PLSA

3.1.1 Introduction

PLSA [3] is a generative model, which can classify data by modeling the data using latent topics. It is a bag of features technique and ignores positional relationship between features in the data.

3.1.2 Problem formulation

The data corpus is of size N. Each document contains features from a code book of size M. Each document is represented as a histogram over the vocabulary. Thus the corpus can be represented by a M by N co-occurrence table, where $n(w_i, d_j)$ stores the number of occurrences of a feature w_i in data example d_j. In addition, there is a hidden (latent) topic variable z_k associated with each occurrence of a feature w_i in an example d_j. The model can be represented as in (Figure 5).

The conditional probability of an observed pair of feature and data example is marginalized over hidden topics and is as follows:

$$P(w_i|d_j) = \sum_{k=1}^{K} P(w_i|z_k)P(z_k|d_j)$$

3.1.3 Model fitting

During model fitting, the task is to evalulate the mixture coefficients $P(w_i|z_k)$ and $P(z_k|d_j)$ $\forall z \in$ Topics, $\forall d \in$ Dataset and $\forall w \in$ Features.

Figure 3: Interest points in s sequence

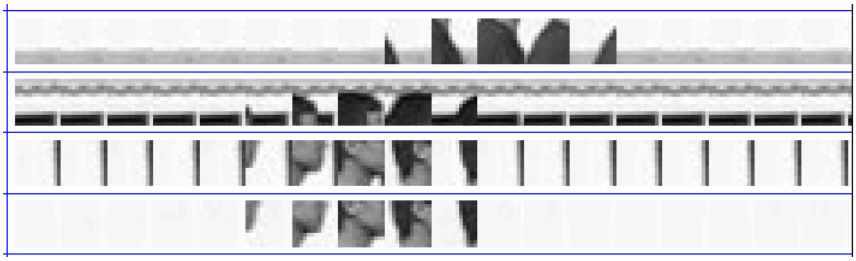

Figure 4: Cuboids

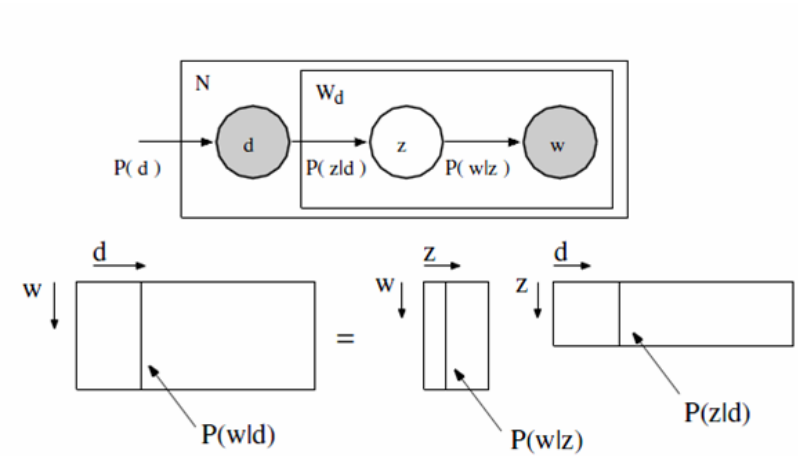

Figure 5: Schematic representation of PLSA

This is achieved by maximizing the liklihood function:

$$\prod_{i=1}^{M}\prod_{j=1}^{N} P(w_i|d_j)^{n(w_i|d_j)}$$

The coefficients are then calculated by using the expectation maximization algorithm. The formulas for the E-step and the M-step have been derived by Hoffman and are reproduced below.

- E-step:

$$P(z_k|d_i, w_j) = \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{l=1}^{K} P(w_j|z_l)P(z_l|d_i)}$$

- M-step:

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i|w_j)P(z_k|d_i, w_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i|w_m)P(z_k|d_i, w_m)}$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i|w_j)P(z_k|d_i, w_j)}{n(d_i)}$$

3.1.4 Unsupervised classification of data examples based on topic

When the task is to cluster data(N) based on topics(K), the following procedure is adopted:

- The co-occurrence table is generated for the corpus of M documents.
- Model fitting is performed using expectation maximization as outlined in the above section.
- The result of model fitting is $P(z_k|d_i)$ and $P(w_j|z_k)$.
- For each document d_i, the topic with maximum $P(z_k|d_i)$ is determined as the representative topic for the document.

3.2 LDA

3.2.1 Introduction

Similar to PLSA, LDA [2] is a generative probablistic model introuced for collections of discrete data such as document corpora. In the generative model, each data example is generated by choosing a distribution over topics and then choosing each word in the document from a topic selected according to the distribution. Various inference techniques can then be used for bayesian parameter estimation which account for the particular data collection. LDA is characterized by assuming dirichlet prior distributions over the documents thus providing a mechanism for generating new documents, thus simplyfying the parameter inference and making it robust against overfitting.

3.2.2 Problem formulation

Frequently, the problem is formulated as follows. The data corpus is D. Each data example contains features from a code book W. The hidden set of topics is T. The model can be represented as in (Figure 6). The probability of an observed feature becomes:

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j)$$

Here, $P(w|z)$ can be represented with T multinomial distributions ϕ over the W code book features, such that $P(w|z = j) = \phi_w^j$. $P(z)$ can be represented with a set of D multinomial distributions θ over the T topics, such that for a feature in a data example d, $P(z = j) = \theta_j^d$. In order

to discover the set of topics used in D, we must estimate ϕ that gives a high probability for the features that occur in the corpus.
In LDA, a prior probability distribution on θ is assumed to provide a generative model for the data examples. Thus ϕ can be estimated without requiring the estimation of θ. In [2], θ and ϕ are not represented as parameters to be estimated. Instead, the posterior distribution over the assignment of features to topics $P(z|w)$ is evaluated. Thus, the probabiliy model for LDA is used with the added Dirichlet prior on ϕ. The complete model then is:

- $w_i|z_i, \phi^{z_i} \sim$ Discrete (ϕ^{z_i})
- $\phi \sim$ Dirichlet (β)
- $z_i|\theta^{d_i} \sim$ Discrete (θ^{d_i})
- $\theta \sim$ Dirichlet (α)

where α and β are hyperparameters specifying the nature of the priors on θ and ϕ. The full model is shown in figure 6.

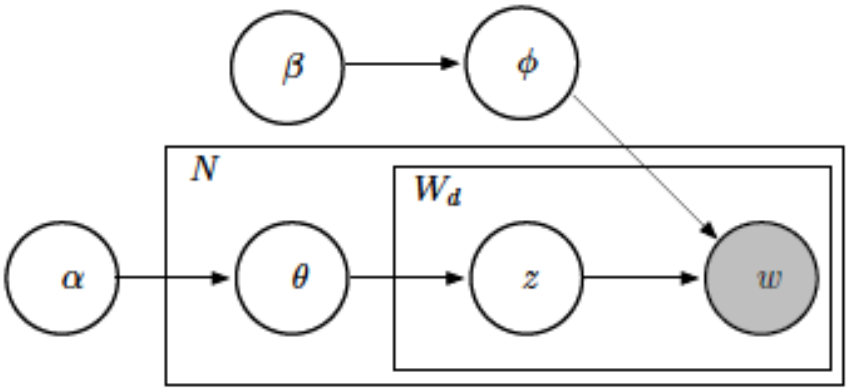

Figure 6: Schematic representation of LDA

3.2.3 Model fitting

Griffiths and Steyvers [2] describe the model fitting process. They propose estimating the posterior $P(z|w)$ using Markov Chain Monte Carlo sampling procedure (Gibbs sampling) inorder to compute the joint probability distribution $p(w, z)$, over a very large discrete space. In the process, two quantities are computed, ${n_j}^d$ which gives the number of times a word in document d has been assigned to topic j and ${n_j}^w$, which gives the number of times a word w has been assigned to topic j. Markov Chain Monte Carlo (Gibbs sampling) is shown below:

$$P(z_i = j|z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

3.2.4 Unsupervised classification of documents based on topic

When the task is to cluster documents(N) based on topics(K), the following procedure is adopted:

- The co-occurrence table is generated for the corpus of M documents.
- Model fitting is performed using Gibbs sampling as outlined in the above section.
- Model fitting provides us with ${n_j}^d$ which gives the number of times a word in document d has been assigned to topic j and ${n_j}^w$ which gives the number of times a word w has been assigned to topic j.
- For each document d, the topic with maximum ${n_j}^d$ is determined as the representative topic for the document.

4 Our Evaluations

4.1 Dataset

We have generated a dataset containing 4 categories of objects in their characteristic motion. The categories are walking humans, moving cars, crawling babies and running cheetahs. Each semantic category contains 10 videos. All videos are approximately 3 seconds shot at 30 fps. The first 2 categories were generated by us around the Dartmouth campus using a video camera. Since we were able to control the data capture process for these categories, we have ensured the stationary nature of the recording camera. The latter 2 category videos were obtained from YouTube. These videos are of a poorer quality compared to the former 2 categories. Also, the camera cannot be deemed as stationary for these videos.
Some sample frames for the various categories are shown below in Figure 7.

4.2 Experiments

We extracted spatio temporal features using the interest point detector implementation provided by the authors of [1]. PCA was used to compress the features to a dimensionality of 100. k-means was used to construct a codebook of size 512. Each video was then represented with a histogram over the codebook. The histograms were not normalized. PLSA and LDA were used for evaluation of their ability to classify the dataset. k-means was also run on the same dataset as a comparison baseline. In all cases, K=4 was used as the number of categories.

4.3 Results

The results are shown in figure 8. As is evident, LDA and PLSA perform very similar and outperform k-means. The results indicate that the task of object classification based on spatio-temporal features can be successfully accomplished using bag of feature learning techniques.

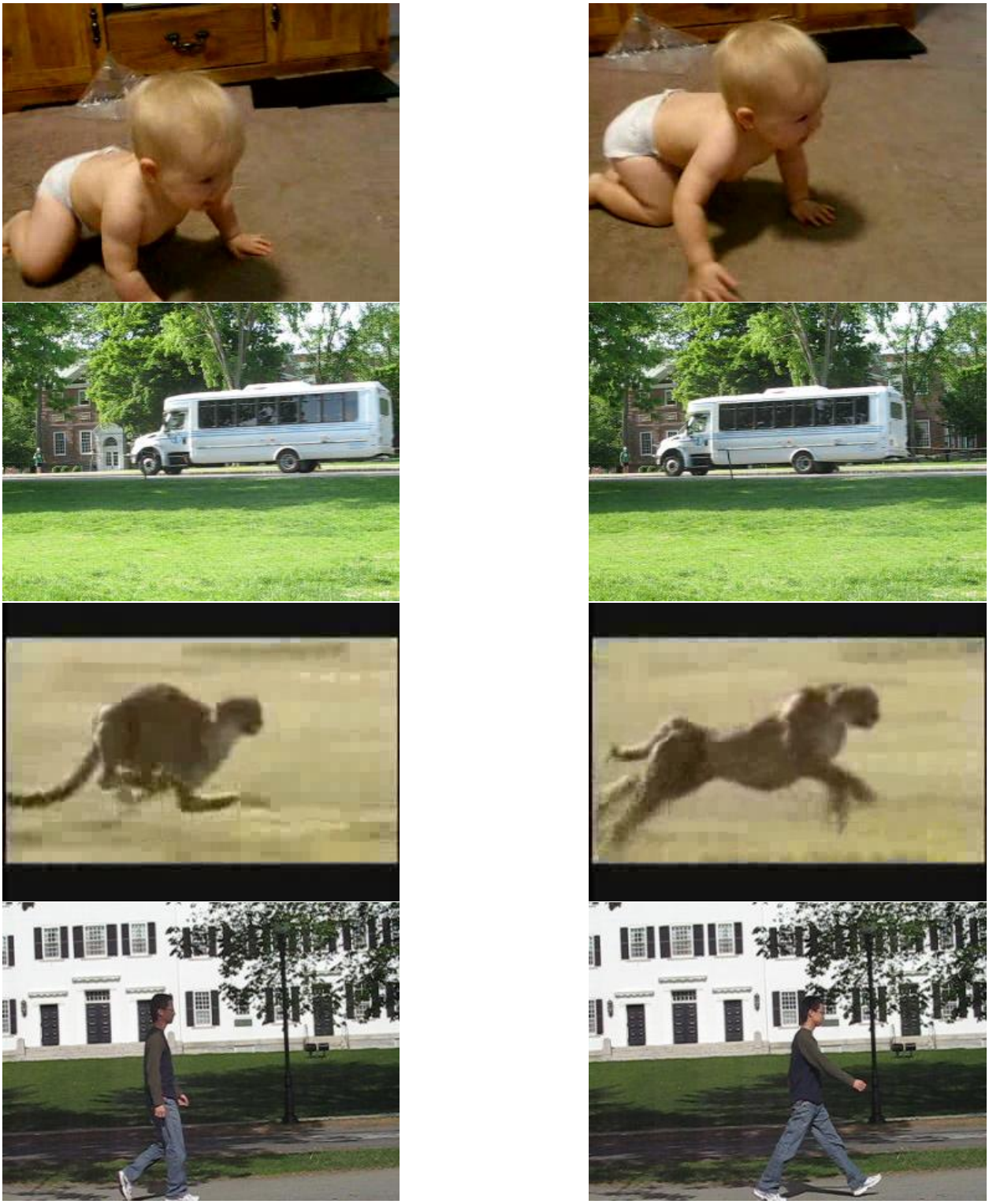

Figure 7: Sample frames from dataset

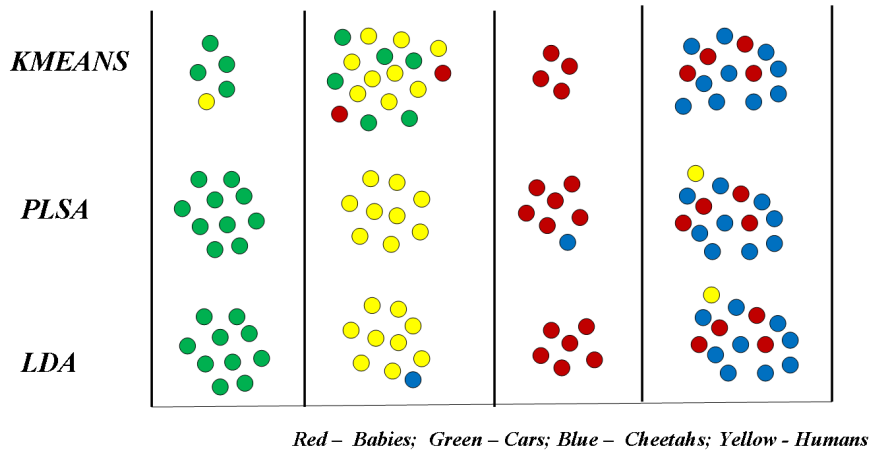

Figure 8: Classification results

References

[1] Piotr Dollár, Vincent Rabaud, Garrison Cottrell and Serge Belongie. Behavior Recognition via Sparse Spatio-Temporal Features. In *Proc. of ICCV VS-PETS*. 2005.

[2] T. Griffiths, M Steyvers. Finding Scientific Topics. In *Proc. Natl. Acad. Sci.*, 2004.

[3] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proc. of SIGIR*, 1999.