

Hidden Markov Models:

Introduction and applications to computer vision

Presented by Omri Guttman

Presentation overview

Intuitive introduction to HMMs:

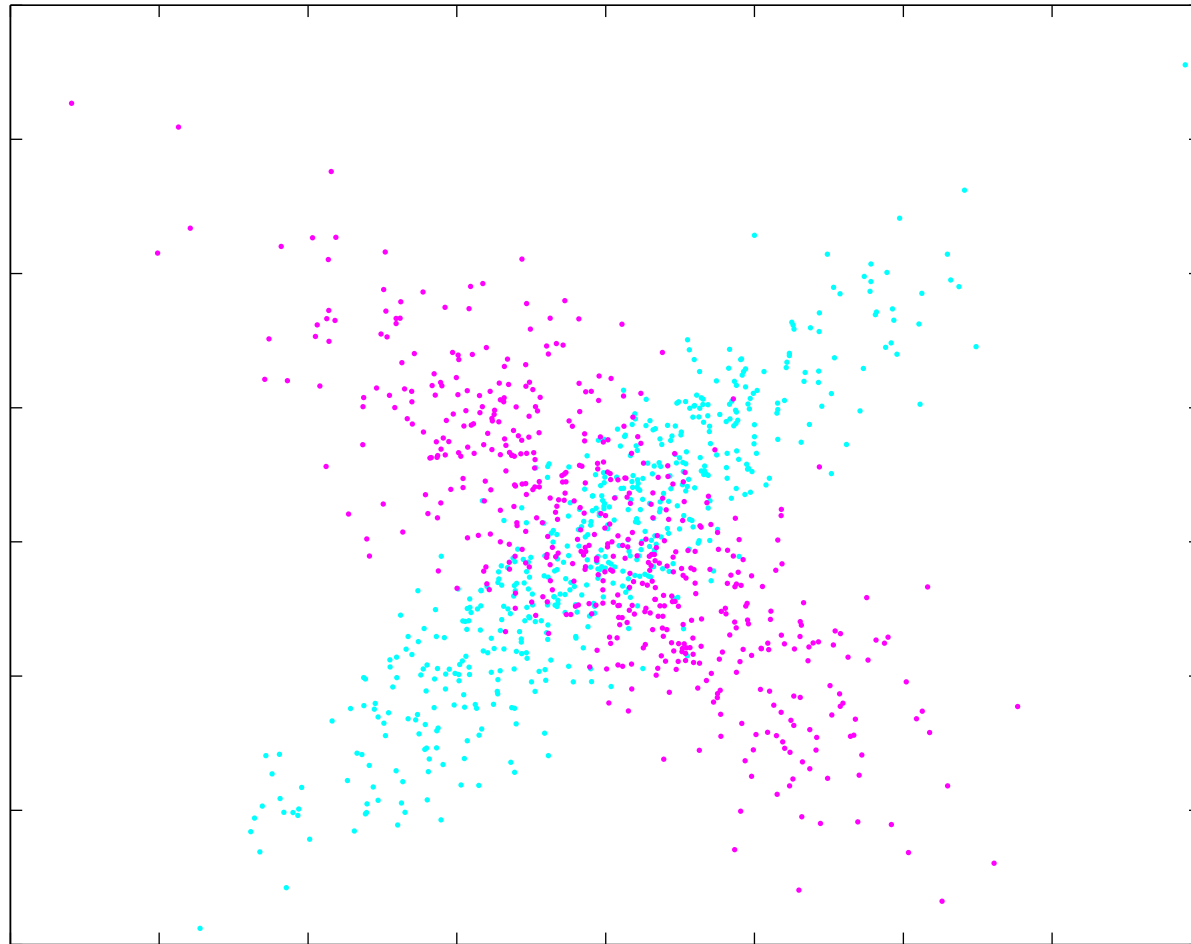
- Modeling static structure.
- Modeling dynamic (temporal) structure.
- Estimating hidden states (decoding).
- Estimating model parameters (learning).
- Connections to different models.

Presentation overview (cont')

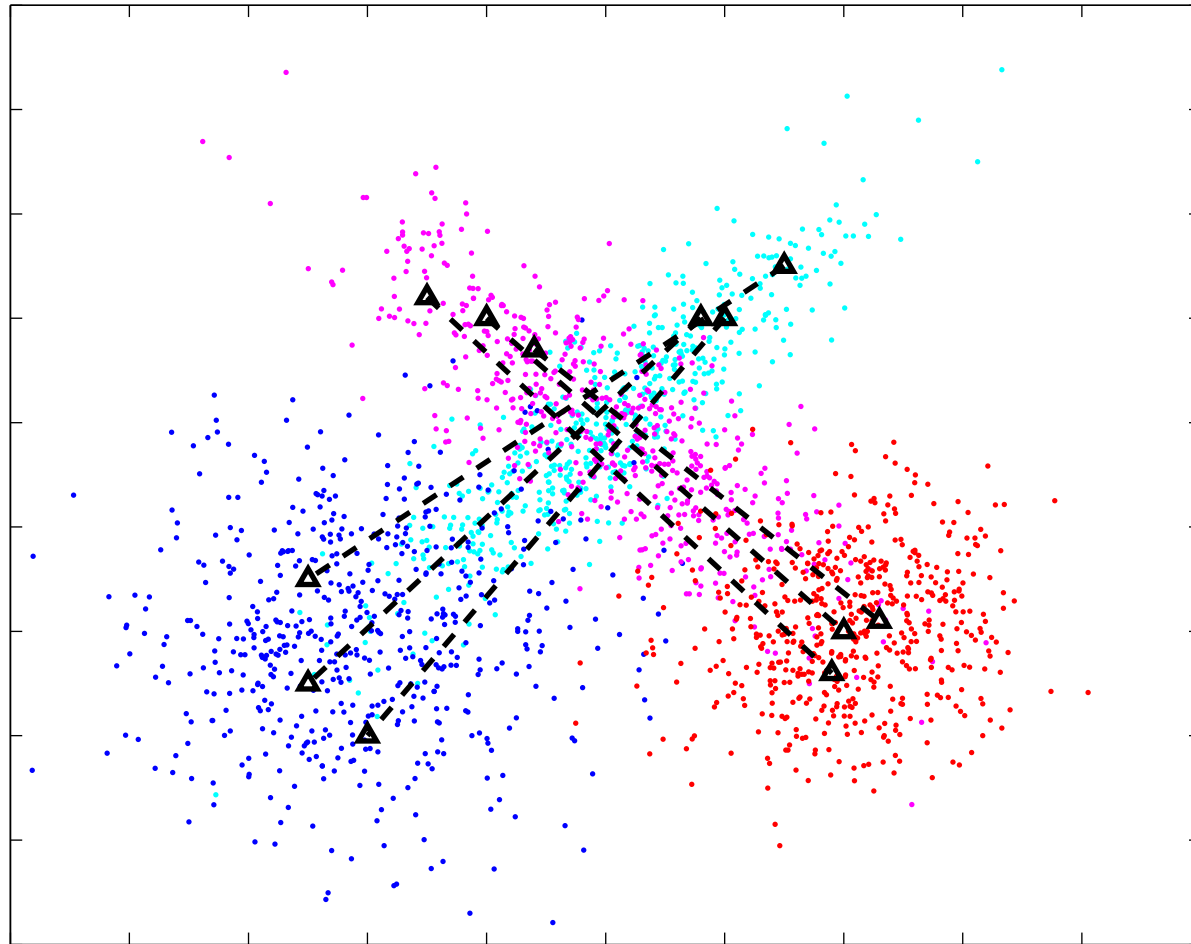
Applications to computer vision:

- Sign language understanding [2,3,4].
- Matching of pictorial structures [5].

Modeling static structure

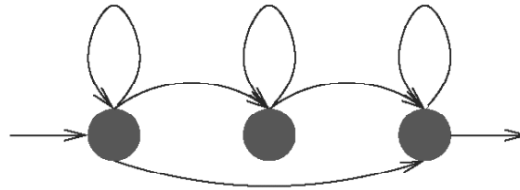


Modeling dynamic (temporal) structure



Graphical model representation

A HMM representing a phoneme can be graphically depicted as follows:



- Each state corresponds to a specific parameterized density model.
- The phoneme is modeled as a 3 state auditory production sequence.
- This topology is typical for modeling processes over time.

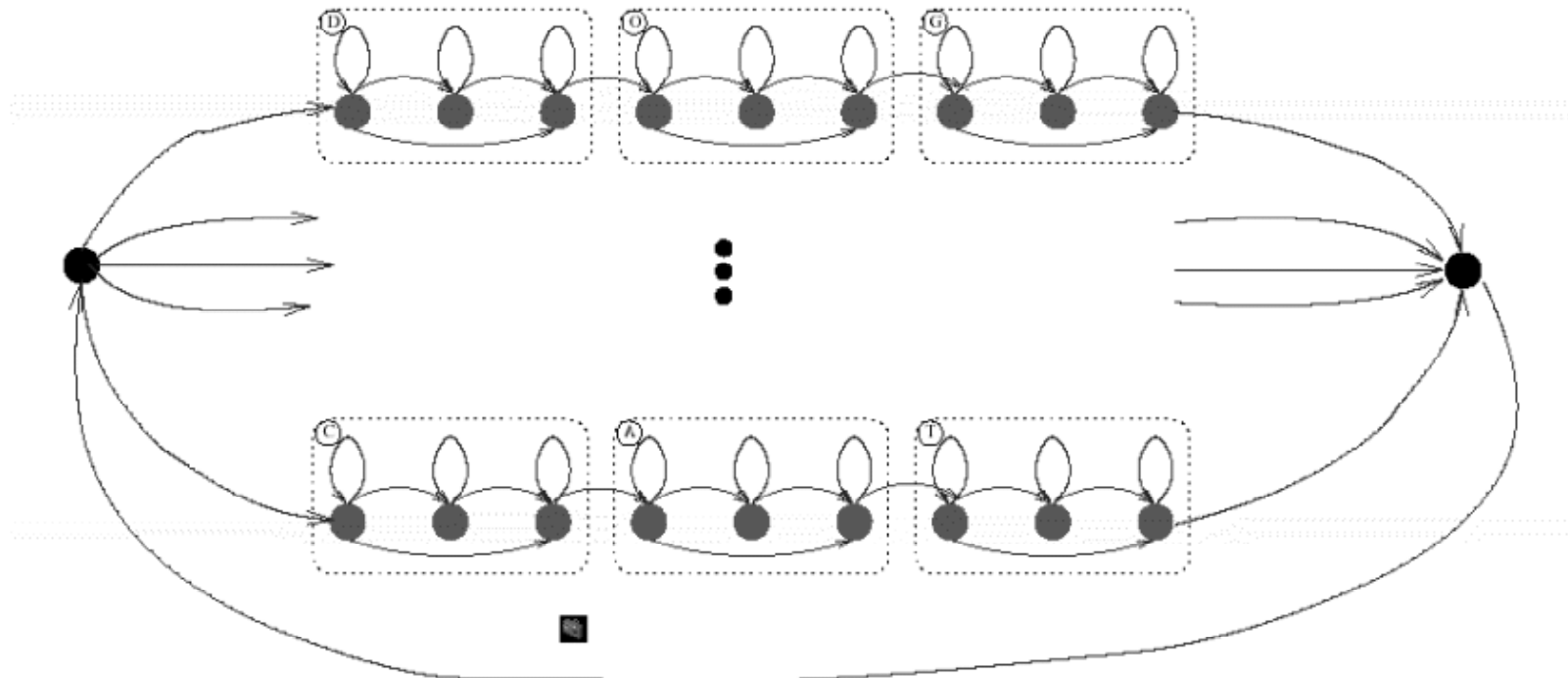
Model assumptions

- Model uses finite number of discrete hidden states.
- First order Markov property:
$$P(q_t | q_{t-1}, q_{t-2}, q_{t-3}, \dots) = P(q_t | q_{t-1})$$

(or $q_{t+1} \perp q_{t-1} | q_t$).
- Emission probability depends only on current state.
- For homogeneous HMMs, all probabilities are time-invariant.

Application to speech recognition

A typical word recognition system will model each word in the vocabulary as a sequence of phoneme models:



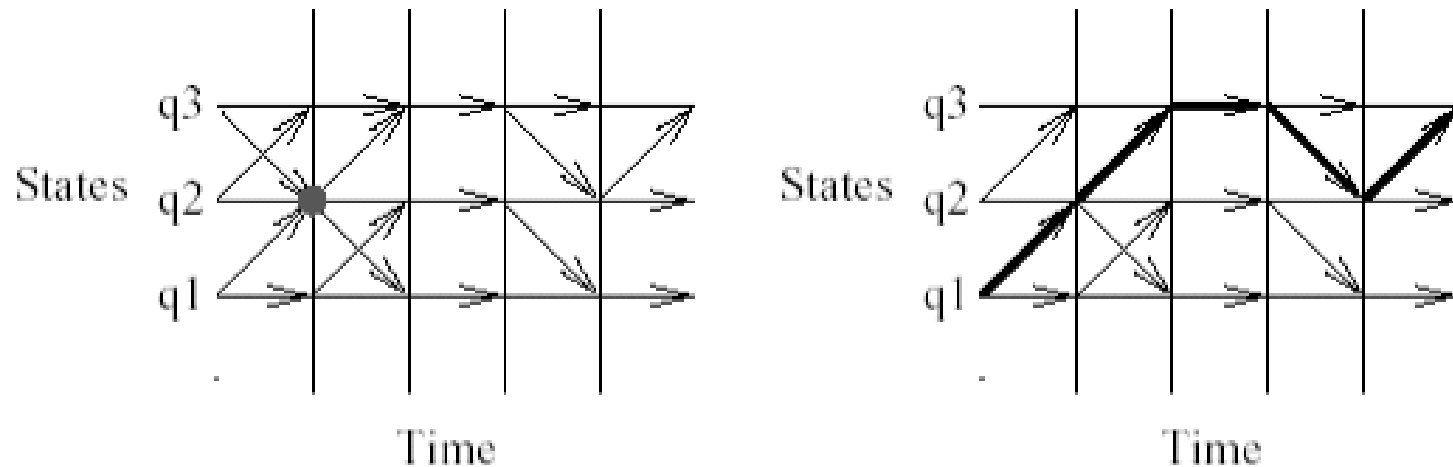
Model parameters

- Initial probability distribution: N parameters.
- State emission probabilities: number of parameters dependent on problem domain and implementation. Typical dimensionality in the speech recognition domain is 30 to 40. Typical parameters from same domain:
 - Discrete codebook (VQ): 128 to 1024 codewords per state.
 - Continuous density: each state modeled as a GMM, with 8 to 64 diagonal covariance Gaussian components.

Model parameters (cont')

- State transition probabilities: $\mathcal{O}(N^2)$ parameters, depending on model structure (typically many transitions are not allowed).
- Total number of parameters in a full blown speech recognition system can easily reach $\mathcal{O}(10^5)$

Estimating hidden states (decoding) Efficient algorithm for inferring the most likely sequence of (hidden) states given a sequence of observations was described by Viterbi, based on dynamic programming, and is of $\mathcal{O}(N^2 \cdot T)$ computational complexity.



Estimating model parameters (learning)

An efficient algorithm for finding a **local** optimum for model parameters for formulated by Baum and Welch, who described EM update equations for the HMM structure.

The B-W algorithm is iterative, with each iteration of $\mathcal{O}(N^2 \cdot T)$ computational complexity.

An intuitive way to think of B-W is as a generalization of fuzzy c-means clustering. A rougher algorithm for learning (often called the Viterbi approximation, corresponding to k-means clustering) which performs hard segmentation is sometimes used for obtaining initial, crude models.

Main advantages of HMMs

- Efficient algorithms for inference and parameter estimation available.
- The often troublesome issue of segmentation is seamlessly integrated into the inference algorithm.

Main shortcomings of HMMs

- Framework is generative (as opposed to discriminative).
- Model structure is typically ad-hoc, and to date no structure learning algorithms have been (well) established in the community.

Connections to different models

- Generalization of Gaussian mixture models.
- Closely related to the Kalman filter.
- Subclass of graphical models. For an enlightening exposition, see [1].

Applications to computer vision

We will examine two applications of Hidden Markov Models to computer vision problems:

- Sign language understanding [2,3,4]. Subproblems in this domain include:
 - Isolated sign recognition.
 - Continuous sign recognition.
- Matching of pictorial structures [5].

Sign language understanding

Many computer vision systems for sign language understanding require specialized equipment (colored gloves, head mounted cameras, camera arrays). Other systems attempt to segment the hands using skin color.

Typical dimensionality is 16, where the extracted features include:

- First and second moments of *blobs* (hand segments), along with time derivatives of above ([4]).
- Three dimensional body part features (position and orientation), extracted using a camera array and a body part identification algorithm ([2,3]).

Pictorial structure matching

The research described in [5] models a car and the human body by graphical models, and the researchers proceed to train and test an efficient structure matching algorithm on the problem domain.

Bibliography

1. K. Murphy, *An introduction to graphical models*, available at:
<http://citeseer.nj.nec.com/murphy01introduction.html>
2. C. Vogler, D. Metaxas, *ASL Recognition Based on a Coupling Between HMMs and 3d Motion Analysis*, available at:
<http://www.cis.upenn.edu/~cvogler/research/research.html>
3. C. Vogler, D. Metaxas, *Parallel Hidden Markov Models for American Sign Language Recognition*, available at:
<http://www.cis.upenn.edu/~cvogler/research/research.html>

4. T. Starner, J. Weaver, A. Pentland, *Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video*, available at:
<http://citeseer.nj.nec.com/starner98realtime.html>

5. P. F. Felzenszwalb, D. P. Huttenlocher, *Efficient Matching of Pictorial Structures*, available at:
<http://www.cs.cornell.edu/~dph/>