

A Factor Graph Framework for Semantic Indexing and Retrieval in Video

Milind R. Naphade Igor Kozintsev, Thomas S. Huang and Kannan Ramchandran
Department of Electrical and Computer Engineering
Beckman Institute for Advanced Science and Technology.
University of Illinois at Urbana-Champaign
{milind,huang,igor,kannan}@ifp.uiuc.edu

Abstract

This paper proposes a novel framework for semantic indexing and retrieval in digital video. The components of the framework are probabilistic multimedia objects (multijects) and a network of such objects (multinets). The main contribution of this paper is a novel application of a factor graph framework to model the interactions in a network of multijects (multinet) at a semantic level. Factor graphs are statistical graphical models that provide an efficient framework for exact and approximate inference via the sum-product algorithm. Incorporating the statistical interactions between the concepts using factor graphs enhances the detection probability of individual multijects and provides a unified framework for integrating multiple modalities and supports inference of unobservable concepts based on their relation with observable concepts. Our experiments reveal significant performance improvement using the inference on the factor graph models.

1. Introduction

The most prominent research in video indexing and retrieval has traditionally focussed on the paradigm of query-by-example (QBE) [12], [11], [2]. Query by keywords (QBK) or key-phrases (preferably semantic) instead of examples has motivated recent research in semantic video indexing. For this we need models which capture the representation corresponding to these keywords. A QBK system can support semantic retrieval for a predetermined set of keywords and also act as the first step in QBE systems to generate examples, some of which can be selected for further queries. The difficulty in supporting semantics lies in the gap between low-level media features and high-level semantics. Recent attempts to address this include the work by Naphade et al [7] which presents hierarchical hidden Markov models for modeling the audio-visual signatures of events like *explosion* and Chang et al [1] who propose a modification to

QBE using semantic visual templates.

In this paper we propose a statistical factor graph framework which attempts to bridge the gap between low-level features and semantic concepts and supports semantic indexing and retrieval. It can also support multiple modalities and fusion of features at various hierarchical levels. Most importantly, for the first time it takes into account the interaction between concepts at various levels of hierarchy. It does this by modeling the interaction between different concepts and providing a mechanism to perform indexing of concepts not observed directly in media. We thus convert the indexing problem into a multimedia pattern recognition problem. Factor graphs come as an elegant model to represent the stochastic relationship between concepts while the sum-product algorithm provides an efficient tool to perform learning and inference for the global model. Our results indicate the promise of our framework which is flexible and universal and provides a good basis for the future extensions and improvements.

2 Proposed Framework

Users of video search engines are interested in retrieving video clips using high-level concepts like *Explosion in a train*. While such semantic queries are very difficult to support exhaustively, they might be supported partially if a model for the event *explosion* exists. User queries might similarly involve concepts like *sky, beach, car* etc. Detection of some of these concepts may be possible, while some others may not be directly observable. To support such queries, we proposed a probabilistic multimedia object (*multiject*) [7] which has a semantic label and which summarizes a time sequence of a set of features extracted from multiple media. *Multijects* can belong to any of the three categories: objects (*car, man, helicopter*), sites (*outdoor, beach*), or events (*explosion, man-walking*). The main contributions of this paper is a framework which explicitly accounts for dependencies between semantic concepts. Intuitively it is clear that the presence of certain multijects suggests a high possibility of

detection of certain other multijects. Similarly some multijects are less likely to occur in the presence of others. The detection of *beach* boosts the chances of *water*, and reduces the chances of detecting *Indoor*. It might also be possible to detect some concepts and infer more complex concepts based on the detected ones. Detection of human speech in the audio stream and a face in the video stream may lead to the inference of *human talking*. To integrate all the multijects and model their interaction, we propose the multinet or network of multijects. A conceptual figure of a multinet is shown in Figure 1 with positive (negative) signs indicating positive (negative) interaction.

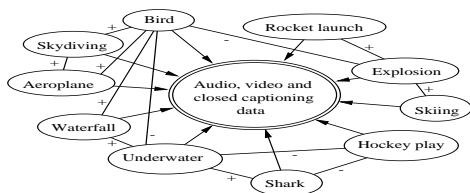


Figure 1. A conceptual figure of a multinet.

3 Video segmentation and Feature Extraction

We now describe the preprocessing steps of segmentation, tracking and feature extraction. Unlike global analysis in [7] we perform our analysis on the dominant regions in video clips. For this we need to segment the video spatio-temporally. The video clips are segmented into shots using the algorithm in [8]. We then use the spatio-temporal segmentation in [12] for each shot to obtain regions homogeneous in color and motion. Dominant regions are then tracked within the shot and labeled. All regions are labeled by a single person choosing from a list of semantic labels. Each region is then processed to extract a set of features which characterize the visual properties including the color, texture, motion, shape and edginess of each region. We extract the following set of features:

Color A linearized 3 channel *HSV* histogram is used, with 12 channels each for *H*, *S* and *V*.

Texture Gray-level co-occurrence matrices (GLCM) of the *V* channel are computed using 32 gray-levels and at 4 orientations: horizontal, vertical, diagonal and anti-diagonal. From these four matrices, six statistical features, contrast, energy, entropy, homogeneity, correlation and inverse difference moment [5] are computed.

Edginess To capture the edges within each region, a Sobel operator with a 3×3 window is applied to each region and the edge map is obtained. Using this edge map an 18 bin histogram of edge directions is obtained as in [4].

Shape Moment invariants as in Dudani et al [3] are used to describe shape of each region.

Motion The inter-frame affine motion parameters for each

region tracked by the spatio-temporal segmentation algorithm are used as motion features.

The feature vector has 84 components for sites (only color, texture and edge features) and 98 components for objects and events.¹ Audio features are extracted as in [9].

4 Modeling semantic concepts using Multijects

We use an identical approach to model concepts in video and audio (independently and jointly). The following site multijects are used in our experiments: *sky*, *water*, *forest*, *rocks* and *snow*. (Results on audio-based multijects like *human-speech*, *music* are presented in [9] and those on audio-visual multijects like *explosion* are presented in [7]). Denoting the feature vector for the region *j* as \vec{X}_j , we model the concept as a binary random variable and define the two hypotheses H_0 and H_1 as

$$H_0 : \vec{X}_j \sim P_0(\vec{X}_j) \quad (1)$$

$$H_1 : \vec{X}_j \sim P_1(\vec{X}_j)$$

where $P_0(\vec{X}_j)$ and $P_1(\vec{X}_j)$ denote the probability density functions conditioned on the null hypothesis (concept absent) and the true hypothesis (concept present). These conditional probability density functions are modeled using a mixture of Gaussian components for the *site* multijects. For *objects* and *events* (in video and audio), hidden Markov models replace the Gaussian mixture models. Feature vectors for all the frames within a shot constitute to the time series modeled by the HMM². Using regions from 18000 frames for training and 94000 frames for testing, the detection performance for the five *site* multijects is given in Table 1.

multiject	Rocks	Sky	Snow	Water	Forest
Detection (%)	77	81.8	81.5	79.4	85.1
False Alarm (%)	24.1	11.9	12.9	15.6	14.9

Table 1. Maximum likelihood classification performance for *site* multijects.

4.1 Frame level semantic features

Since the multijects are used as semantic feature detectors at a regional level, it is easy to define features at the frame level by integrating the region-level classification. One problem is that if more than one concept lies in a single segment

¹Automatically finding the most important features for any classification task is a problem in itself, which we do not address here.

²The models for hypothesis H_1 used 5 gaussian components while those for H_0 used 10. The number of mixture components was fixed experimentally and could be different for optimal performance. In general models for H_0 are represented better with more components than those for H_1

and if the segment is classified as belonging to one of the five multijects, we are losing information. To avoid this we check each segment for each concept individually and obtain probabilities of each concept being present or absent in the segment. Imperfect segmentation thus does not hurt us since mistakes made here can be incorporated as soft decisions and probabilistically corrected in the multinet. Let R_{ij} denote a binary random variable which takes value 1 when concept i is present in region j and value 0 otherwise. Assuming uniform priors on the presence or absence of a given concept in any region and using the Bayes' rule we then obtain:

$$P(R_{ij} = 1 | \vec{X}_j) = \frac{P(\vec{X}_j | R_{ij} = 1)}{P(\vec{X}_j | R_{ij} = 1) + P(\vec{X}_j | R_{ij} = 0)} \quad (2)$$

A concept is present in the image (frame) if it is present in any of the regions. Let us define binary random variables F_i , $i \in \{1, N\}$ (N is the number of concepts) which takes on a value of 1 if concept i is present in the current frame and takes the value 0 otherwise. The operation which combines information for each concept from all regions to give information about F_i is the *OR* operation. Using the compact notation $\mathcal{X} = \{\vec{X}_1, \dots, \vec{X}_M\}$ (M is the number of regions in a frame) the conditional probabilities for presence and absence of concepts at the frame level are computed as

$$P(F_i = 0 | \mathcal{X}) = \prod_{j=1}^{j=M} P(R_{ij} = 0 | \vec{X}_j) \quad (3)$$

$$P(F_i = 1 | \mathcal{X}) = 1 - P(F_i = 0 | \mathcal{X})$$

5 Factor graphs for multijects

To model the interaction between multijects in a multinet, we propose to use a novel *factor graph* [6] framework. Factor graphs were initially successfully applied in the area of channel error correction coding [6] and, specifically, iterative decoding. Turbo decoding and other iterative decoding techniques have in the last few years proven to be landmark developments in coding theory. The unsurpassed performance-complexity tradeoff of iteratively decodable codes has led to an explosion of work in this area. Before explaining how factor graphs can be used to build a multinet we introduce some necessary notation. Let $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ be a vector of variables. A *factor graph* visualizes the factorization of a global function $f(\mathbf{x})$. Let $f(\mathbf{x})$ factor³ as $f(\mathbf{x}) = \prod_{i=1}^m f_i(\mathbf{x}^{(i)})$, where $\mathbf{x}^{(i)}$ is the set of variables of the function f_i . A factor graph for f is defined as the bipartite graph with two vertex classes V_f and V_v of sizes m and n respectively such that the i th node in V_f is connected to the j th node in V_v iff f_i is a function of x_j . The variables

³The definition of multiplication and addition may be taken rather generally. For the precise algebraic properties that multiplication and addition have to satisfy, see [6].

may be from different alphabets as long as the functions f_i are properly defined.

Many problems in signal processing and learning are formulated as minimizing or maximizing a global function $f(\mathbf{x})$ marginalized for a subset of its arguments. The algorithm which allows us to perform this efficiently, though in most cases only approximately, is called the **sum-product algorithm**.

5.1 The sum-product algorithm in a tree

In this section we provide an example of the sum-product algorithm applied to a finite tree. A factor graph is a *tree* if and only if there exists a unique path between any two nodes. Suppose we are given the following factorization of the probability density function defined over six variables⁴:

$$f(x_1, x_2, x_3, s_1, s_2, s_3) = f_1(s_1)f_2(s_1, x_1)f_3(s_1, s_2)f_4(s_2, x_2)f_5(s_2, s_3)f_6(s_3, x_3)$$

Suppose now that we are interested in the *a posteriori* distribution of s_1 after observing the values x'_1, x'_2, x'_3 of x_1, x_2, x_3 :

$$P(s_1, x_1 = x'_1, x_2 = x'_2, x_3 = x'_3) = \sum_{s_2, s_3} f(x_1 = x'_1, x_2 = x'_2, x_3 = x'_3, s_1, s_2, s_3)$$

These computations may be performed efficiently in a distributed manner using the following sum factorization⁵:

$$P(s_1, x_1 = x'_1, x_2 = x'_2, x_3 = x'_3) = f_1(s_1)f_2(s_1, x_1 = x'_1) \left\{ \sum_{s_2} f_3(s_1, s_2)f_4(s_2, x_2 = x'_2) \right\} \left\{ \sum_{s_3} f_5(s_2, s_3)f_6(s_3, x_3 = x'_3) \right\} \quad (4)$$

In Equation 4 two types of computations are performed: multiplication of local functions and marginalization with respect to local variables. First, f_5 and f_6 are multiplied producing the result which depends on s_2 and s_3 . Then this result is summed over s_3 to produce the function of s_2 only. The process repeats for other variables. We can think of the operations just described in terms of *passing messages* from all nodes of the graph along the edges according to some *schedule*. The messages are formed using the following simple rules: a message from a function node to a variable node is the product of all messages incoming to the function node with the function itself, marginalized for the variable associated with the variable node; a message from a variable node

⁴This factorization may represent a simple Hidden Markov Chain if s_1, s_2, s_3 are the hidden states and x_1, x_2, x_3 are observations. Function f_1 is the prior distribution of the initial state, functions f_3 and f_5 are conditional distributions of the next state given the present state, functions f_2, f_4 and f_6 are conditional distributions of observations given the state.

⁵For an HMM interpretation this is the well-known forward-backward algorithm for HMMs [10] and is an instance of the **sum-product** algorithm.

to a function node is simply the product of all messages incoming to the variable node from other functions connected to it. Further details about the sum-product algorithm are found in [6].

If a factor graph is not a tree, the sum-product algorithm can still be used to perform an approximate marginalization. Surprisingly, it is exactly the situation, where the sum-product algorithm shows an excellent performance in many signal processing applications. This is also a situation of interest for the problem of semantic indexing of video since the relations between concepts are complicated and, in general, contain numerous cycles (e.g., see Figure 1).

5.2 Global factor graph

Our approach to video indexing is based on constructing a global graphical model which describes the probabilistic relations between various multijects in the frame. We use the five site multijects *forest*, *sky*, *snow*, *rocks* and *water* to classify each region and obtain five likelihoods ratios for each region of the form $\frac{Prob(feature|observation)}{Prob(no\ feature|observation)} = \frac{P(R_{ij}=1|X_j)}{P(R_{ij}=0|X_j)}$. The likelihoods for the features are fed into the deterministic OR function to produce the likelihoods at the frame level as explained before. Additionally, a global joint distribution connects the features in the frame. The factor graph in Figure 2 represents the factorization of the global probability density function defined over the set of local and global features and observation vectors.

In general, the global distribution function defines a length- 2^N probability mass vector. While this is obviously the most general case, in practice, it may be undesirable to estimate all 2^N entries. Alternatively, we can only allow nonzero two-factor terms in the global function (Each function has two variables as its arguments and there are C_2^N such terms) thus reducing the complexity. This approach is illustrated in Figure 3. The potential problem in this case may be the fact that the exact inference becomes hard as the number of variables grows (due to the loops in the graph). We then apply iterative techniques based on the sum-product algorithm to overcome this.

The proposed factor graph framework provides an efficient way to represent probabilistic and deterministic relations between concepts. In such scenarios the sum-product algorithm provides us with a computationally efficient tool for learning and inference.

6 Experimental Setup

We have digitized video data from movies of different genres. For the site multijects we use as many as 18000 frames for training and 94000 frames for testing. We perform classification for each concept on each region and then merge results at the frame-level using the OR function. The regions in the training set images and the corresponding

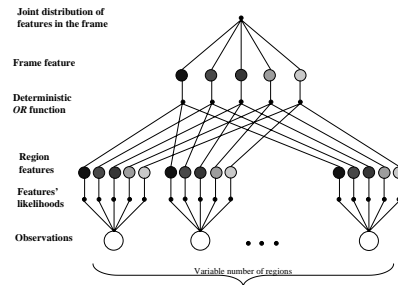


Figure 2. Image segments (regions) are processed independently by using *multijects* and decisions are fused to obtain frame-level features. Additionally, a global joint distribution connects the features in the frame.

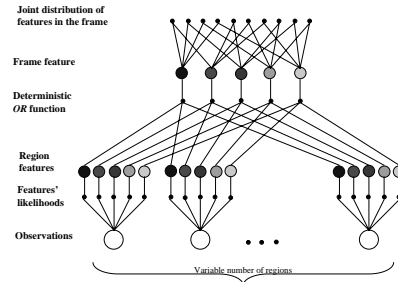


Figure 3. A factor graph for a frame with the joint distribution of features in the frame represented by two-factors only.

ground truth are used to train the multijects for the five concepts. The decisions at region level are merged at the frame level to obtain variables, F_i .

7 Results

To evaluate the results we compare the detection performance of the multijects with and without accounting for the concept dependencies. The reference system performs multiject detection by thresholding soft-decisions (i.e., $P(F_i|\mathcal{X})$) at the frame-level. The proposed system performs detection by thresholding the feature distributions obtain using both the observations and the joint distribution function of the features in the frame (which is obtained as a result of iterative probability propagation in the factor graphs). We use receiver operating characteristics (ROC) curves which show a plot of the probability of detection plotted against the probability of false alarms for different values of a parameter (the threshold in our case).

Figure 4 shows the ROC curves for the overall performance across all the five multijects. The three curves in the ROC correspond to the performance using the reference frame-level classification, a factor graph shown in Figure 2, and a factor graph shown in Figure 3. Note, that in the last case the factor graph is not a tree and we perform an approximate inference using the sum-product algorithm⁶. Clearly,

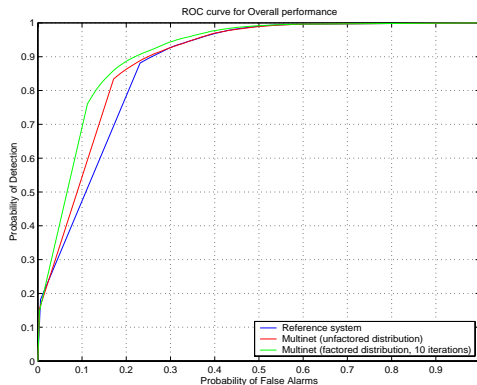


Figure 4. ROC curves for overall performance

there is significant improvement in detection performance by using the multinet than without using it. This improvement is especially stark for low P_f . In fact the improvement is more than 22 % for a threshold corresponding to $P_f = 0.1$. What is even more interesting is that the detector based on the factorized function (Figure 3) performs better than the one based on the unfactored function. This may suggest that the factorized function is a better representative for the concept dependencies than the one shown in Figure 2 due to the fact that the factorized function is estimated more reliably (it has less coefficients to estimate).

8 Conclusions and Future Research

We propose a statistical factor graph framework for detection of semantic concepts using multiple media. We show that it is beneficial to account for the interaction between semantic concepts. To this end we propose and implement a multinet which enhances the detection accuracy for the concepts significantly. We also demonstrate the power of iterative inference technique based on the sum-product algorithm which can handle complex dependencies and cycles in the graph. The multinet does not impose any conditions on the choice of the classifiers used for detecting the concepts independently in the first phase and we can improve performance by selecting better classifiers in the primary stage. Future research includes demonstrating the ability of the factor graph multinet to seamlessly integrate multiple modalities at various levels and to support inference of concepts (that are not

⁶We set the number of iterations to be 10 which we found to be sufficient from our experiments. The ROC curves are evaluated at 2000 threshold values between 0 and 1

directly observed in terms of media features) through their interaction with those concepts which are modeled using representations in media features. Future research also aims at modeling the temporal dynamics of the relationships using dynamic multinets.

9 Acknowledgments

Milind Naphade was supported in part by a fellowship from the Computational Science and Engineering Department at the University of Illinois and in part by NSF Grant CDA 96-24396. Igor Kozintsev was supported in part by NSF Grant CCR 99-79381. The authors would like to thank D. Zhong and S. F. Chang for the spatio-temporal segmentation algorithm.

References

- [1] S. F. Chang, W. Chen, and H. Sundaram. Semantic visual templates - linking features to semantics. In *Proceedings of the fifth IEEE International Conference on Image Processing*, volume 3, pages 531–535, Chicago, IL, Oct 1998.
- [2] Y. Deng and B. S. Manjunath. Content based search of video using color, texture and motion. In *Proceedings of the IEEE International Conference on Image Processing*, volume 2, pages 13–16, Santa Barbara, CA, Oct 1997.
- [3] S. Dudani, K. Breeding, and R. McGhee. Aircraft identification by moment invariants. *IEEE Trans. on Computers*, C-26(1):39–45, Jan 1977.
- [4] A. K. Jain and A. Vailaya. Shape-based retrieval: A case study with trademark image databases. *Pattern Recognition*, 31(9):1369–1390, 1998.
- [5] R. Jain, R. Kasturi, and B. Schunck. *Machine Vision*. MIT Press and McGraw-Hill, 1995.
- [6] F. Kschischang, B. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *submitted to IEEE Trans. Inform. Theory*, July 1998.
- [7] M. Naphade, T. Kristjansson, B. Frey, and T. S. Huang. Probabilistic multimedia objects (multijects): A novel approach to indexing and retrieval in multimedia systems. In *Proceedings of the fifth IEEE International Conference on Image Processing*, volume 3, pages 536–540, Chicago, IL, Oct 1998.
- [8] M. Naphade, R. Mehrotra, A. M. Ferman, J. Warnick, T. S. Huang, and A. M. Tekalp. A high performance shot boundary detection algorithm using multiple cues. In *Proceedings of the fifth IEEE International Conference on Image Processing*, volume 2, pages 884–887, Chicago, IL, Oct 1998.
- [9] M. R. Naphade and T. S. Huang. Stochastic modeling of soundtrack for efficient segmentation and indexing of video. In *SPIE IS & T Storage and Retrieval for Multimedia Databases*, volume 3972, pages 168–176, Jan 2000.
- [10] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, Feb 1989.
- [11] H. Zhang, A. Wang, and Y. Altunbasak. Content-based video retrieval and compression: A unified solution. In *Proceedings of the IEEE International Conference on Image Processing*, volume 1, pages 13–16, Santa Barbara, CA, Oct. 1997.
- [12] D. Zhong and S. F. Chang. Spatio-temporal video search using the object-based video representation. In *Proceedings of the IEEE International Conference on Image Processing*, volume 2, pages 21–24, Santa Barbara, CA, Oct. 1997.