Visual Cue Cluster Construction via Information Bottleneck Principle and Kernel Density Estimation

Winston H. Hsu and Shih-Fu Chang

Dept. of Electrical Engineering, Columbia University, New York, NY 10027, USA
{winston, sfchang}@ee.columbia.edu

Abstract. Recent research in video analysis has shown a promising direction, in which mid-level features (e.g., people, anchor, indoor) are abstracted from low-level features (e.g., color, texture, motion, etc.) and used for discriminative classification of semantic labels. However, in most systems, such mid-level features are selected manually. In this paper, we propose an information-theoretic framework, visual cue cluster construction (VC^3), to automatically discover adequate mid-level features. The problem is posed as mutual information maximization, through which optimal cue clusters are discovered to preserve the highest information about the semantic labels. We extend the Information Bottleneck framework to high-dimensional continuous features and further propose a projection method to map each video into probabilistic memberships over all the cue clusters. The biggest advantage of the proposed approach is to remove the dependence on the manual process in choosing the mid-level features and the huge labor cost involved in annotating the training corpus for training the detector of each mid-level feature. The proposed VC^3 framework is general and effective, leading to exciting potential in solving other problems of semantic video analysis. When tested in news video story segmentation, the proposed approach achieves promising performance gain over representations derived from conventional clustering techniques and even the mid-level features selected manually.

1 Introduction

In the research of video retrieval and analysis, a new interesting direction is to introduce “mid-level” features that can help bridge the gap between low-level features and semantic concepts. Examples of such mid-level features include location (indoor), people (male), production (anchor), etc., and some promising performance due to such mid-level representations have been shown in recent work of news segmentation and retrieval [1, 2]. It is conjectured that mid-level features are able to abstract the cues from the raw features, typically with much higher dimensions, and provide improved power in discriminating video content of different semantic classes. However, selection of the mid-level features is typically manually done relying on expert knowledge of the application domain. Once the mid-level features are chosen, additional extensive manual efforts are needed to annotate training data for learning the detector of each mid-level feature.

Our goal is to automate the selection process of the mid-level features given defined semantic class labels. Given a collection of data, each consisting of low-level features and associated semantic labels, we want to discover the mid-level features automatically. There is still a need for labeling the semantic label of each data sample, but the large cost associated with annotating the training corpus for each manually chosen mid-level feature is no longer necessary. In addition, dimensionality of the mid-level features will be much lower than that of the low-level features.

Discovery of compact representations of low-level features can be achieved by conventional clustering methods, such as K-means and its variants. However, conventional methods aim at clusters that have high similarities in the low-level feature space but often do not have strong correlation with the semantic labels. Some clustering techniques, such as LVQ [3], take into account the available class labels to influence the construction of the clusters and the associated cluster centers. However, the objective of preserving the maximum information about the semantic class labels was not optimized.

Recently, a promising theoretic framework, called Information Bottleneck (IB), has been developed and applied to show significant performance gain in text categorization [4, 5]. The idea is to use the information-theoretic optimization methodology to discover "cue word clusters" which can be used to represent each document at a mid level, from which each document can be classified to distinct categories. The cue clusters are the optimal mid-level clusters that preserve the most of the mutual information between the clusters and the class labels.

In this paper, we propose new algorithms to extend the IB framework to the visual domain, specifically video. Starting with the raw features such as color, texture, and motion of each shot, our goal is to discover the cue clusters that have the highest mutual information about the final class labels, such as video story boundary or semantic concepts. Our work addresses several unique challenges. First, the raw visual features are continuous (unlike the word counts in the text domain) and of high dimensions. We propose a method to approximate the joint probability of features and labels using kernel density estimation. Second, we propose an efficient sequential method to construct the optimal clusters and a merging method to determine the adequate number of clusters. Finally, we develop a rigorous analytic framework to project new video data to the visual cue clusters. The probabilities of such projections over the cue clusters are then used for the final discriminative classification of the semantic labels.

Our work is significantly different from [6] which uses the IB principle for image clustering. In [6], 3 CIE-Lab colors and 2 horizontal and vertical positions are used as the input raw features. The dimension is much lower than that in this paper. The distribution in the raw feature space was first fit by a Gaussian Mixture Model (GMM), whose estimated parameters were then used for the IB clustering. In contrast, we do not assume specific parametric models in our approach, making our results more generalizable. Most importantly, preservation of mutual information about the semantic labels was not addressed in [6].

We test the proposed framework and methods in story segmentation of news video using the corpus from TRECVID 2004 [7]. The results demonstrate that when combined with SVM, projecting videos to probabilistic memberships among the visual cue clusters is more effective than other representations such as K-means or even the manually selected mid-level features. An earlier unoptimized implementation was submitted to TRECVID 2004 story segmentation evaluation and achieved a performance very close to the top.

The main idea of the IB principle and its extension to high-dimensional continuous random variables are introduced in Section 2. The discriminative model and the feature selection based on the induced VC^3 clusters are presented in Section 3. In Section 4, evaluation of the proposed techniques in news video story segmentation is described. We present conclusions and future work in Section 5.

2 The Information Bottleneck Principle

The variable X represents (feature) objects and Y is the variable of interest or auxiliary labels associated with X. X might be documents or low-level feature vectors; Y might be document types in document categorization or sematic class labels. In this context, we want the mapping from $x \in X$ to cluster $c \in C$ to preserve as much information about Y as possible. As in the compression model, the framework passes the information that X provides about Y through a “bottleneck” formed by the compact summaries in C. On the other hand, C is to catch the consistent semantics of object X. The semantic is defined by the conditional distribution over the auxiliary label Y.

Such goal can be formulated by the IB principle, which states that among all the possible clusterings of the objects into a fixed number of clusters, the desired clustering is the one that minimizes the loss of mutual information (MI) between the features X and the auxiliary labels Y. Assume that we have joint probability $p(x, y)$ between these two random variables. According to the IB principle, we seek a clustering representation C such that, given a constrain on the clustering quality $I(X;C)$, the information loss $I(X,Y) - I(C;Y)$ is minimized.

2.1 Mutual Information

For discrete-valued random variables X and Y, the MI between them is $I(X;Y) = \sum_y \sum_x p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$. We usually use MI to measure the dependence between variables. In the VC^3 framework, we represent the continuous D-dimensional features with random variable $X \in R^D$; the auxiliary label is a discrete-valued random variable Y representing the target labels. We have feature observations with corresponding labels in the training set $S = \{x_i, y_i\}_{i=1..N}$. Since X is continuous, the MI is defined as $I(X;Y) = \sum_y \int_x p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dx$. However, based on S, the practical estimation of MI from the previous equation is difficult. To address this problem, the histogram approach is frequently used but only works between two scalars. An alternative approach is to model X through

GMM which is limited to low-dimensional features due to the sparsity of data in high-dimensional spaces.

We approximate the continuous MI with Eq. 1 for efficiency. The summarization is only over the observed data x_i assuming that $p(x,y) = 0$ if $x \notin S$. Similar assumptions are used in other work (e.g., the approximation of Kullback-Leibler divergence in [6]). According to our experiments, the approximation is satisfactory in measuring the MI between the continuous feature variable X and the discrete auxiliary variable Y.

$$I(X;Y) \cong \sum_i \sum_y p(x_i,y) \log \frac{p(x_i,y)}{p(x_i)p(y)} \tag{1}$$

2.2 Kernel Density Estimation

To approximate the joint probability $p(x,y)$ based on the limited observations S, we adopt the kernel density estimation (KDE) [8]. The method does not impose any assumption on the data and is a good method to provide statistical modeling among sparse or high-dimensional data.

The joint probability $p(x,y)$ between the feature space X and the auxiliary label Y is calculated as follows:

$$p(x,y) = \frac{1}{Z(x,y)} \sum_{x_i \in S} K_\sigma(x - x_i) \cdot \bar{p}(y|x_i), \tag{2}$$

where $Z(x,y)$ is a normalization factor to ensure $\sum_{x,y} p(x,y) = 1$, K_σ (Eq. 3) is the kernel function over the continuous random variable X. $\bar{p}(y|x_i)$ is an un-smoothed conditional probability of the auxiliary labels as observing feature vector x_i. We assume that Y is binary in this experiment and $\bar{p}(y|x_i)$ is either 0 or 1. Note that Y can extend to multinomial cases in other applications.

From our observation, $\bar{p}(y|x_i)$ is usually sparse. Eq. 2 approximates the joint probability $p(x,y)$ by taking into account the labels of the observed features but weighted and smoothed with the Gaussian kernel, which measures the non-linear kernel distance from the feature x to each observation x_i. Intuitively, nearby features in the kernel space will contribute more to Eq. 2.

Gaussian kernel K_σ for D-dimensional features is defined as:

$$K_\sigma(x_r - x_i) = \prod_{j=1}^{D} \exp \frac{-||x_r^{(j)} - x_i^{(j)}||}{\sigma_j}, \tag{3}$$

where $\sigma = [\sigma_1, .., \sigma_j, .., \sigma_D]$ is the bandwidth for kernel density estimation. We can control the width of the bandwidth to embed prior knowledge about the adopted features; for example, we might emphasize more on color features and less on the texture features by changing the corresponding σ_j.

2.3 Sequential IB Clustering

We adopt the sequential IB (sIB) [4] clustering algorithm to find clusters under the IB principle. It is observed that sIB converge faster and is less sensitive to local optima comparing with other IB clustering approaches [4].

The algorithm starts from an initial partition C of the objects in X. The cluster cardinality $|C|$ and the joint probability $p(x,y)$ are required in advance. At each step of the algorithm, one object $x \in X$ is drawn out of its current cluster $c(x)$ into a new singleton cluster. Using a greedy merging criterion, x is assigned or merged into c^* so that $c^* = \operatorname{argmin}_c d_F(\{x\}, c)$. The merging cost, the information loss due to merging of the two clusters, represented as $d_F(c_i, c_j)$, is defined as (cf. [5] for more details):

$$d_F(c_i, c_j) = (p(c_i) + p(c_j)) \cdot D_{JS}[p(y|c_i), p(y|c_j)], \tag{4}$$

where D_{JS} is actually Jensen-Shannon (JS) divergence and $p(c_i)$ and $p(c_j)$ are cluster prior probabilities. JS divergence is non-negative and equals zero if and only if both its arguments are the same and usually relates to the likelihood measure that two samples, independently drawn from two unknown distributions, are actually from the same distribution.

The sIB algorithm stops as the number of new assignments, among all objects X, to new clusters are less than a threshold, which means that so far the clustering results are "stable." Meanwhile, multiple random initialization is used to run sIB multiple times and select the results that has the highest cluster MI $I(C;Y)$, namely the least information loss $I(X;Y) - I(C;Y)$.

2.4 Number of Clusters

To learn the optimal number of clusters in the clustering algorithm is still an open issue. G-means is one of the options but limited to low-dimensional data due to its Gaussian assumption. IB proposes a natural way to determine the number of clusters by discovering the break point of MI loss along the agglomerative IB (aIB) clustering algorithm [5, 6]. The algorithm is a hard clustering approach and starts with the trivial clustering where each cluster consists of a single item. To minimize the overall information loss, a greedy bottom-up process is applied to merge clusters that minimize the criterion in Eq. 4, which states the information loss after merging clusters c_i and c_j. The algorithm ends with a single cluster with all items. Along the merging steps, there is a gradual increase in information loss. We can determine the "adequate" number of the clusters by inspecting the point where a significant information loss occurs.

3 Discriminative Model

3.1 Feature Projection

We use VC^3 to provide a new representation of discriminative features by transforming the raw visual features into the (soft) membership probabilities over those induced cue clusters which have different conditional probability $p(y|c)$ over the auxiliary label Y.

Each key frame with raw feature vector $\mathbf{x_r}$ is projected to the induced clusters and represented in visual cue feature $\mathbf{x_c}$, the vector of membership probabilities over those K induced visual cue clusters;

$$\mathbf{x_c} = [x_c^1, ..., x_c^j, ..., x_c^K], \tag{5}$$

$$x_c^j = \hat{p}(c_j|\mathbf{x_r}) = \frac{J(c_j|\mathbf{x_r})}{\sum_{k=1}^{K} J(c_k|\mathbf{x_r})}, \text{and} \tag{6}$$

$$J(c_j|\mathbf{x_r}) = p(c_j) \cdot \hat{p}(x_r|c_j) = p(c_j) \cdot \frac{1}{|c_j|} \sum_{\mathbf{x_i} \in c_j} K_\sigma(\mathbf{x_r} - \mathbf{x_i}). \tag{7}$$

$J(c_j|\mathbf{x_r})$ is proportional to the (soft) posterior probability $\hat{p}(c_j|\mathbf{x_r})$ depicting the possibility that the raw feature $\mathbf{x_r}$ belongs to cluster c_j, hence, can be represented by the product of the cluster prior $p(c_j)$ and the cluster likelihood $\hat{p}(\mathbf{x_r}|c_j)$; the latter is also estimated with KDE based on the visual features within the cluster c_j. The visual cue features $\mathbf{x_c}$ is later used as the input feature for discriminative classification. With this feature projection, we represent the raw feature $\mathbf{x_r}$ with the membership probabilities towards those visual cue clusters. Each cluster has its own semantic defined by the auxiliary label Y since all the visual features clustered into the same cluster have similar condition probability over Y.

3.2 Support Vector Machines

SVM has been shown to be a powerful technique for discriminative learning [9]. It focuses on structural risk minimization by maximizing the decision margin. We applied SVM using the Radial Basis Function (RBF) as the kernel, $K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2), \gamma > 0$.

In the training process, it is crucial to find the right parameters C (tradeoff on non-separable samples) and γ in RBF. We apply five fold cross validation with a grid search by varying (C, γ) on the training set to find the best parameters achieving the highest accuracy.

3.3 Feature Selection

After sIB clustering, the cluster MI between the induced feature clusters C and auxiliary label Y is measured with $I(C;Y) = \sum_c I(c)$ and can be decomposed into summation of the MI contribution of each cluster c, defined in Eq. 8. We further utilize this property to select the most significant clusters with the highest $I(c)$, on the other hand, to remove less significant or unstable clusters.

$$I(c) \equiv p(c) \sum_y p(y|c) \log \frac{p(c, y)}{p(c)p(y)} \tag{8}$$

4 Experiments

4.1 Broadcast News Story Segmentation

We tested the proposed VC^3 approach on the story segmentation task in TRECVID [7]. A news story is defined as a segment of news broadcast with a coherent news focus which contains at least two independent declarative clauses. Story boundary detection is an interesting and challenging problem since there are no simple fixed rules of productions or features [10].

To solve this problem, researchers try different ways to manually enumerate the important production cues, and then train the specific classifiers to classify them. For example, in [1], 17 domain-specific detectors are manually selected and trained. In [11], a large set of manually picked features are fused using statistical methods like maximum entropy.

4.2 Approach: Discriminative Classification

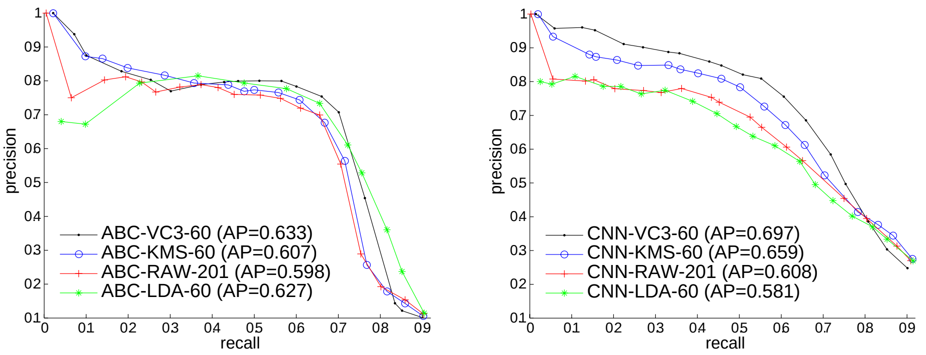
We train a SVM classifier to classify a candidate point as a story boundary or non-boundary. The major features for the discriminative model is the visual cue features represented in the membership probabilities (Section 3.1) towards the induced visual cue clusters. Applying the VC^3 framework, the continuous random variable X now represents the concatenated raw visual features of 144-dimensional color autocorrelogram, 9-dimensional color moments, and 48-dimensional Gabor textures for each key frame (See explanations in [7]). The label Y is binary, "story" and "non-story."

The number of visual cue clusters is determined by observing the break point of accumulated MI loss as described in Section 2.4 and is 60 both for ABC and CNN videos. To induce the visual cue clusters, 15 videos for each channel are used; 30 videos, with key frames represented in the cue cluster features, for each channel are reserved for SVM training; the validation set is composed of 22 CNN and 22 ABC videos. They are all from TRECVID 2004 development set.

4.3 Results and Discussions

We present the boundary detection performance in terms of the precision and recall (PR) curve and Average Precision (AP) which averages the (interpolated) precisions at certain recalls. For a $M+1$ point AP, $AP = \frac{1}{M+1}\sum_{i=0}^{M} P(r_i)$; $r_i = i/M$ indicates the designated recall sample; $P(r_i) = \max_{r_i \le r} P(r)$ is the interpolated precision, where $\{r, P(r)\}$ are those available recall-precision pairs from the classification results. Intuitively, AP can characterize the PR curve in a scalar. A better classifier, with a PR curve staying upper-right corner of the PR plane, will have higher AP, and vice versa. In this experiment, we set $M = 20$.

Fig. 1(a) and 1(b) show the discriminative classification of story boundaries on ABC and CNN videos in PR curves and APs. All boundary detection use SVM but on different feature configurations. The VC^3 approach on both video sets (ABC/CNN-VC3-60) performs better that those with raw visual features



(a)

(b)

Fig. 1. (a): PR curves of story boundary detection on ABC videos with feature configurations via VC^3 (ABC-VC3-60), K-means (ABC-KMS-60), raw visual features (ABC-RAW-201), and LDA (ABC-LDA-60); (b) the same as (a) but on CNN videos. The corresponding AP of each PR curve is shown as well.

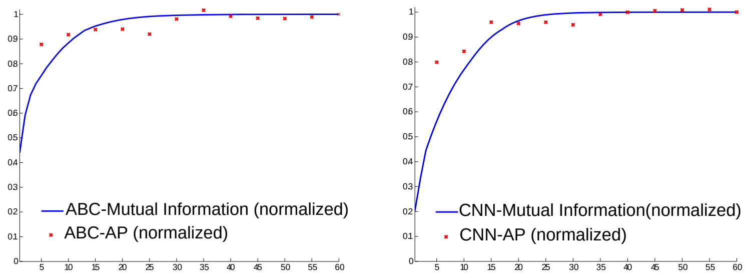
(ABC/CNN-RAW-201). The performance gap is significant in CNN due to the diversity of the channel's production effects. Those semantic representations from mid-level cue clusters benefit the boundary classification results.

With VC^3, we transform 201-dimensional raw visual features into 60-dimensional semantic representations. To show the effectiveness of this approach, we compare that with the feature reduction via Linear Discriminative Analysis (LDA), which usually refers to a discriminative feature transform that is optimal when the data within each class is Gaussian [12]. LDA features (ABC-LDA-60) perform almost the same with VC^3 in ABC and even better than those raw visual features, but not in CNN videos. It is understandable because diversity of CNN breaks the Gaussian assumption of LDA.

Comparing with the K-means[1] approach (ABC/CNN-KMS-60), which clusters features considering only Euclidean distance in the feature space, the VC^3 discriminative features (ABC/CNN-VC3-60) perform better in both channels. The reason is that VC^3 clustering takes into account the auxiliary (target) label rather than by feature similarity only, which is what K-means is restricted to.

Even with the same cluster number, the cluster MI $I(C;Y)$ through VC^3 is larger than that through K-means; e.g., $I(C;Y)$ is 0.0212 for VC^3 and 0.0193 for K-means in ABC, and 0.0108 and 0.0084 respectively in CNN. The difference between K-means and VC^3 MI in CNN videos is more significant than that in ABC videos. It might explain why CNN VC^3 has more performance gain

[1] For fair comparison, "soft" membership probability of Eq. 6 is used to derive features towards those K-means clusters and significantly outperforms the common "hard" membership.



(a) (b)

Fig. 2. Relation of preserved MI and AP of top N visual clusters; (a) normalized AP vs. MI of the top N selected visual cue clusters in ABC. (b) AP vs. MI in CNN.

over the K-means approach. In ABC, since positive data mostly form compact clusters in the feature space (e.g., boundaries are highly correlated with anchors, etc.), the VC^3 does not differ a lot from other approaches.

4.4 Feature Selection

In feature selection among those induced visual cue clusters, the accumulated MI between the top N visual cue clusters (x-axis), $\sum_{i=1}^{N} I(c_i)$, and detection AP, are shown in Fig. 2. The MI curves and classification performance (AP) are all normalized by dividing the corresponding values with all (60) cue clusters. The results show that preserved MI of the selected cue clusters is a good indicator of the classification performance. It also allows us to determine the required number of clusters by applying a lower threshold to the cumulative MI. As seen in Fig. 2(b), CNN videos need more cue clusters to reach the same AP.

4.5 VC^3 vs. Prior Work

Evaluated on the same CNN validation set, the VC^3 approach described in this paper, with automatically induced visual features **only**, has AP=0.697. When augmented with speech prosody features, the performance improves to 0.805 AP and outperforms our previous work [10], which fuses detectors of anchors, commercials, and prosody-related features through SVM (AP=0.740) on the same data set. More discussions regarding multi-modality fusion and their performance breakdowns in different (visual) story types can be seen in [7].

5 Conclusion and Future Work

We have proposed an information-theoretic VC^3 framework, based on the Information Bottleneck principle, to associate continuous high-dimensional visual features with discrete target labels. We utilize VC^3 to provide new representation for discriminative classification, feature selection, and prune "non-informative" visual feature clusters. The proposed techniques are general and effective, achieving close to the best performance in TRECVID 2004 story segmentation. Most importantly, the framework avoids the manual procedures to select features and greatly reduces the amount of annotation in the training data.

Some extensions of VC^3 to induce audio cue clusters, support multi-modal news tracking and search are under investigation. Other theoretic properties such as automatic bandwidth selection for KDE and performance optimization are also being studied.

Acknowledgments

We thank Dan Ellis and Lyndon Kennedy of Columbia University for useful discussions and V. France of [12] for his kind support of LDA implementation. This material is based upon work funded in whole by the U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.

References

1. Chaisorn, L., Chua, T.S., , Koh, C.K., Zhao, Y., Xu, H., Feng, H., Tian, Q.: A two-level multi-modal approach for story segmentation of large news video corpus. In: TRECVID Workshop, Washington DC (2003)
2. Amir, A., Berg, M., Chang, S.F., Iyengar, G., Lin, C.Y., Natsev, A., Neti, C., Nock, H., Naphade, M., Hsu, W., Smith, J.R., Tseng, B., Wu, Y., Zhang, D.: IBM research trecvid 2003 video retrieval system. In: TRECVID 2003 Workshop. (2003)
3. Kohonen, T.: Self-Organizing Maps. third edn. Springer, Berlin (2001)
4. Slonim, N., Friedman, N., Tishby, N.: Unsupervised document classification using sequential information maximization. In: 25th ACM international Conference on Research and Development of Information Retireval. (2002)
5. Slonim, N., Tishby, N.: Agglomerative information bottleneck. In: Neural Information Processing Systems (NIPS). (1999)
6. Gordon, S., Greenspan, H., Goldberger, J.: Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations. In: International Conference on Computer Vision. (2003)
7. Hsu, W., Kennedy, L., Chang, S.F., Franz, M., Smith, J.: Columbia-IBM news video story segmentation in trecvid 2004. (Technical Report ADVENT #207-2005-3)
8. Scott, D.W.: Multivariate Density Estimation : Theory, Practice, and Visualization. Wiley-Interscience (1992)
9. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
10. Hsu, W., Chang, S.F.: Generative, discriminative, and ensemble learning on multi-modal perceptual fusion toward news video story segmentation. In: IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan (2004)
11. Hsu, W., Chang, S.F., Huang, C.W., Kennedy, L., Lin, C.Y., Iyengar, G.: Discovery and fusion of salient multi-modal features towards news story segmentation. In: IS&T/SPIE Electronic Imaging, San Jose, CA (2004)
12. France, V., Hlavac, V.: Statistical pattern recognition toolbox for matlab. Technical report, Czech Technical University (2004)