KEY-FRAME EXTRACTION FOR OBJECT-BASED VIDEO SEGMENTATION

Xiaomu Song and Guoliang Fan
School of Electrical and Computer Engineering
Oklahoma State University
Stillwater, OK 74078

ABSTRACT

We propose an coherent approach to extract key-frames within a video shot for object-based video segmentation. A unified feature space is first constructed to represent video frames and visual objects simultaneously in a joint spatio-temporal domain, and key-frame extraction is formulated as a feature selection process that aims to maximize the cluster divergence of video objects by selecting an optimal set of key-frames. Specifically, two different criteria are used to achieve joint key-frame extraction and object segmentation. One criterion recommends key-frame extraction that leads to the maximum pairwise interclass divergence between objects in the feature space. The other aims at maximizing the marginal divergence of objects in each frame. Simulations with both synthetic and real video data manifest the efficiency and robustness of the proposed methods.

1. INTRODUCTION

Object-based video segmentation is a fundamental step towards content-based video analysis. Recently, a statistical model-based segmentation method was developed to coherently segment video objects in a joint spatio-temporal domain [1]. In this method, all frames in a video shot are considered as one entity for the model estimation that supports object-based segmentation. In order to approximate the nonlinear nature of motion patterns, the work in [2] suggests to split a video shot into a succession of block of frames (BOF) with certain overlaps, and the model estimation and object segmentation are performed within each BOF individually, where the motion could be approximately linear. All video frames or BOFs may contain considerably redundant information regarding statistical modeling of video objects in the feature space. Outliers, such as noise and insignificant objects that might randomly appear in a video shot, could also cause more overlaps between clusters. These factors may degrade the accuracy of the model estimation, and thus deteriorate the performance of object segmentation.

This work was supported by the National Science Foundation (NSF) under Grant IIS-0347613 (CAREER) and the DEPSCoR under Grant W911NF-04-1-0221.

In order to improve the efficiency and robustness of statistical model-based object segmentation, we recently proposed a so-called *combined key-frame extraction and object-based video segmentation* approach in [3], where the model estimation is conducted based on a set of pre-selected key-frames. Compared with the methods in [1, 2], this approach can significantly reduce the computational load and also enhance the performance of object segmentation. However, the method in [3] separates key-frame extraction and object segmentation as two sequential steps, where the relationship between key-frames and objects is not revealed. This may still lead to a redundant feature space where the overlapping problem cannot be mitigated effectively.

In this work, we propose a coherent approach to extract video key-frames for object segmentation. A unified feature space is first constructed to represent video frames and objects coherently in the joint spatio-temporal domain. Then key-frame extraction is formulated as a feature selection problem that aims at maximizing the cluster divergence in the feature space. Specifically, two criteria are used for key-frame extraction to optimize object segmentation. One is the maximum average interclass Kullback Leibler distance (MAIKLD), the other is the maximum marginal divergence (MMD) [4]. MAIKLD considers both temporal and spatial correlations between frames, and requires combinatorial search of key-frames. MMD tries to maximize the variance to the mean density of all object classes in each frame individually, so that it can be implemented more efficiently than MAIKLD. Compared with MAIKLD, MMD may generate less representative key-frames for object segmentation. Generally, the proposed methods can provide compact and salient key-frame sets that support robust object segmentation.

2. UNIFIED FEATURE SPACE

Video key-frame extraction and object segmentation are usually based on different feature subsets. A unified feature subset is needed for coherent key-frame extraction and object segmentation. This feature subset should contain both spatial and temporal information. In this work, we use a pixel-wise 7-D feature vector suggested in [3], including

(Y, u, v) color features, spatial location (x, y), time t, as well as *intensity change over the time* to provide additional motion information.

Generally, the frames within a video shot represent a spatially and temporally continuous action, and share the common visual and often semantic-related characteristics, leading to tremendously redundant information in the feature space. As mentioned before, outliers could cause more overlaps between major objects (clusters). One example is shown in Fig. 1, where a video shot of N frames contains three objects. Since we fix the feature dimension, we want to reduce the overlapping problem by extracting a set of key-frames that can effectively support object representation in the feature space. Then the model estimation and object segmentation can be efficiently accomplished based on these key-frames.

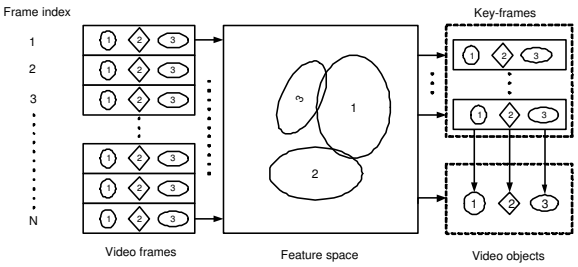


Fig. 1. A unified feature space.

Even though object segmentation is usually an unsupervised process where no prior knowledge is available, we still can extract the most salient and useful key-frames via cluster divergence measurements in the feature space. Thus, in this work, *key-frame extraction is formulated as a feature selection process where key-frames are extracted by maximizing the cluster divergence regarding different objects.*

3. MAXIMUM AVERAGE INTERCLASS KULLBACK LEIBLER DISTANCE

Kullback Leibler distance (KLD) [5] is one often used divergence measurement. Given two probability densities $p_i(x)$ and $p_j(x)$, the KLD between them is defined as:

$$KL(p_i, p_j) = \int p_i(x) \ln \frac{p_i(x)}{p_j(x)} dx, \tag{1}$$

KLD is usually not a symmetric distance measurement and can be symmetrized by adding $KL(p_i, p_j)$ and $KL(p_j, p_i)$ together. Ideally, the larger the KLD, the more separability between clusters. If there are M clusters, the average interclass KLD (AIKLD) is defined as:

$$\bar{D} = C \sum_{i=1}^{M} \sum_{j>i}^{M} [KL(p_i, p_j) + KL(p_j, p_i)], \tag{2}$$

where $C = \frac{2}{M(M-1)}$.

Let $\mathbf{X} = \{\mathbf{x}_i,\ 1 \le i \le N\}$ with cardinality $|\mathbf{X}| = N$ be an original video shot with N frames and M objects. Let $\mathbf{Z} = \{\mathbf{x}_i^*,\ 1 \le i \le N^*\}$ be any subset of $\mathbf{X}$ with cardinality $|\mathbf{Z}| = N^* \le N$. Then key-frame extraction is aimed to find a $\mathbf{X}^*$, which is one of $\mathbf{Z}$, so that $\bar{D}$ can be maximized:

$$\mathbf{X}^* = \arg \max_{\mathbf{Z} \in \mathbf{X}, |\mathbf{Z}| \le N} \bar{D}_{\mathbf{Z}}, \tag{3}$$

where $\bar{D}_{\mathbf{Z}}$ is the AIKLD of M objects within $\mathbf{Z}$ in the 7-D feature space. Maximum AIKLD (MAIKLD) is optimal in the sense of minimum Bayes error [6]. Nevertheless, it is not easy to find an optimal solution, especially when N is large. A computationally efficient suboptimal solution is more preferred in practice. In this work, a combinatorial feature selection method, i.e., Sequential Forward Floating Selection (SFFS) [7], is used to extract video key-frames. In the following, we call the frames to be tested for key-frame extraction *key-frame candidates*.

For simplicity, we do not begin with all frames in $\mathbf{X}$, and apply the method in [8, 3] to extract $N' \le N$ initial key-frame candidates (still redundant), where a similarity measurement based on the framewise 2-D Hue and Saturation (HS) color histogram is used. Then the Gaussian mixture model (GMM) is used to model video objects coherently in the unified feature space based on these key-frame candidates. The iterative Expectation maximization (EM) algorithm [9] is applied with the minimum description length (MDL) model selection criterion [10]. After the model estimation, the objects in all key-frame candidates are segmented out using the maximum *a posteriori* criterion. Then SFFS is applied to extract key-frames that maximize AIKLD of the objects. SFFS is initialized via the Sequential Forward Selection (SFS) method to generate a combination that comprises 2 key-frame candidates. The algorithm is stopped when N^* is greater than or equal to a threshold (e.g., $N'/2$), or the iteration reaches a given number (e.g., 20).

The proposed segmentation method has several significant advantages: (1) It is computationally efficient based on a small set of key-frames. (2) The optimal or near-optimal set of key-frames can be extracted for robust object segmentation. (3) The algorithm is flexible without significant data-dependent thresholds. Nevertheless, SFFS may not be efficient enough when N' is large.

4. MAXIMUM MARGINAL DIVERSITY

In order to further improve the key-frame extraction process at a lower computational complexity, we also suggest another method based on the marginal cluster divergence in each key-frame. In a recent work [4], a maximum marginal diversity (MMD) criterion is proposed for efficient feature

selection with very simple computations. Under certain constraints, MMD is equivalent to the *infomax* principle [11] that is also optimal in the sense of minimum Bayes error. In the context of classification, the *infomax* principle indicates that any feature selection method should select certain features that maximize the mutual information (MI) between the features and class labels [4]. When the *infomax* principle is applied to key-frame extraction and object segmentation, the objective function is defined as:

$$\mathbf{X}^* = \arg \max_{\mathbf{Z} \in \mathbf{X}, |\mathbf{Z}| \leq N} I(\mathbf{Z}, Y), \tag{4}$$

where $\mathbf{X}$, $\mathbf{X}^*$, and $\mathbf{Z}$ are defined as (3), and $I(\mathbf{Z}, Y)$ is the MI between $\mathbf{Z}$ and the class label $Y = \{1, 2, \cdots, M\}$. It was derived in [4] that:

$$\begin{aligned} I(\mathbf{Z}, Y) &= E_Y[KL(p(\mathbf{Z}|Y = y), p(\mathbf{Z}))] \\ &= \sum_{i=1}^{N^*} MD(\mathbf{x}_i^*) + \epsilon, \end{aligned} \tag{5}$$

where

$$\begin{aligned} MD(\mathbf{x}_i^*) &= E_Y[KL(p(\mathbf{x}_i^*|Y = y), p(\mathbf{x}_i^*))], \\ \epsilon &= \sum_{i=2}^{N^*} I(\mathbf{x}_i^*; \mathbf{x}_{1,i-1}^*|Y) - \sum_{i=2}^{N^*} I(\mathbf{x}_i^*; \mathbf{x}_{1,i-1}^*), \end{aligned} \tag{6}$$

and $\mathbf{x}_{1,i-1}^* = \{\mathbf{x}_1^*, \mathbf{x}_2^*, \cdots \mathbf{x}_{i-1}^*\}$. $MD(\mathbf{x}_i^*)$ is called the marginal diversity (MD) [4], which means the variance of the mean density. ϵ shows the information of class labels conveyed in the MI between features.

The analysis in [4] indicates that the solutions of MMD and *infomax* are equal when $\epsilon = 0$, which means the mutual information between features is not affected by class labels. As generalized in [4], this condition is originated from the recent researches about image statistics, which suggest that a rough structure of pattern dependencies between some image features follows general statistical laws that are independent of class labels. Although this might not be always strictly held, at least it proves that MMD could be optimal under such condition.

When implementing MMD, similar to MAIKLD, key-frame extraction is performed after the GMM model estimation. MMD considers the cluster divergence in each key-frame candidate, and extract N^* of them that have the largest MD values. This process considerably mitigates the computational load. N^* could be predetermined, or be adaptively determined given a threshold of the MD value. In the simulation, we set the average MD of all key-frame candidates as the threshold. A key-frame candidate is extracted as the key-frame if its MD is greater than the threshold.

MAIKLD tries to maximize the expectation of the pairwise inter-class divergence, while MMD criterion aims at maximizing the average divergence to the mean density. Accordingly, they lead to different results although both of them could be lower bounded by Bayes error. MAIKLD should result in more representative key-frames regarding video objects than MMD because a large variance of the mean density cannot guarantee good separabilities between clusters. However, MMD is faster than MAIKLD because no combinatorial search is necessary.

5. SIMULATIONS AND DISCUSSIONS

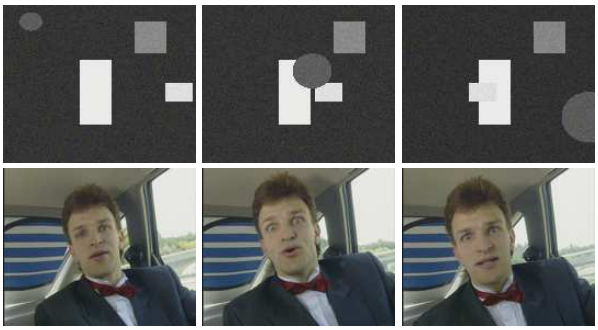
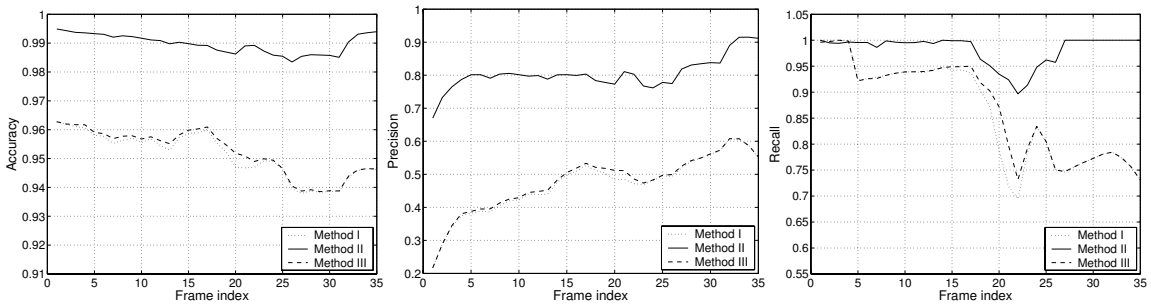


Fig. 2. A synthetic video (the first row) and a real video (the second row). The frame size is 176×144.

Simulations are performed on a computer with 3.2GHz CPU and 1GB memory. The proposed methods are tested on both synthetic and real video sequences as shown in Fig. 2. The purpose of using a synthetic video is to numerically evaluate the object segmentation performance, where we calculate segmentation *accuracy*, *precision*, and *recall* with respect to all moving objects. The first row of Fig. 2 shows three frames in the synthetic video where an elliptic object is moving diagonally from the top-left to the bottom-right corner, and its size is increasing simultaneously. A rectangular object is moving from right to left horizontally. Additionally, some Additive White Gaussian Noise (AWGN) is added to the synthetic video.

We denote the method in [3] as Method-I, and two proposed methods as Method-II (MAIKLD) and Method-III (MMD). The numerical results of the three methods on the synthetic video are shown in Fig. 3. It can be seen that Method-II outperforms Method-I although Method-II uses less key-frames for object segmentation. Method-III uses the same number of key-frames as method-II, but its performance is inferior to that of Method-II. This indicates that Method-II can extract more representative and salient key-frames regarding the video objects than Method-III. Both Methods-I and -III result in low recalls due to the fact that the extracted key-frames do not support accurate model estimation. Method-III slightly outperforms Method-I because Method-III uses spatial information to characterize both key-frames and objects in the feature space.

We also compare three methods on the real video shown in Fig. 2. The number of the initial key-frame candidates and the extracted key-frames are listed in Table 1. It is



(a) Accuarcy (b) Precision (c) Recall

Fig. 3. Numerical results: dotted, solid, and dashed lines indicate results of Methods-I, -II, and -III, respectively.

Table 1. Key-frame numbers (KFN) and computation time (CT).

Video sequences	Method-I		Method-II		Method-III	
	KFN	CT (s)	KFN	CT (s)	KFN	CT (s)
Synthetic (36 frames)	19	169	9	186	9	175
Real (150 frames)	16	187	8	210	9	193

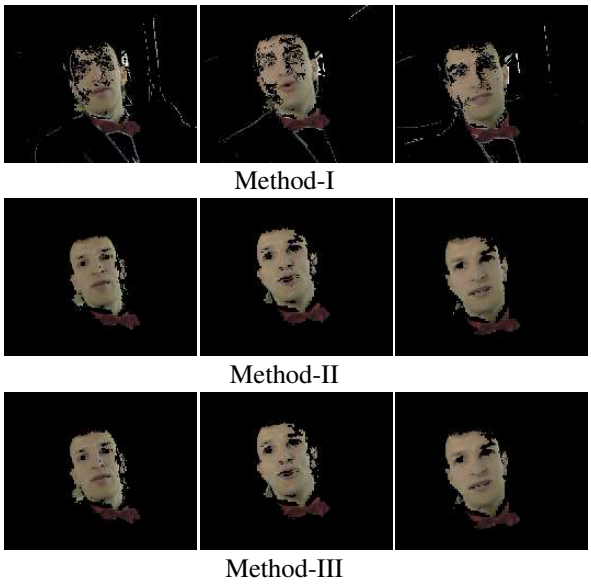


Fig. 4. Segmentation results of the real video based on the same number of key-frames (e.g., 8 key-frames).

shown that Methods-II and -III slightly increase the computational load compared with Method-I, but generate more compact and representative key-frames. In order to compare the three methods in terms of their effectiveness of key-frame extraction for object segmentation, we fix the number of extracted key-frames to be 8. Fig. 4 illustrates the segmentation results of the three methods. It can be seen that Methods-II and III significantly outperform Method-I that extracts the key-frames only using the framewise color histogram. Method-II extracts the key-frames by considering joint spatio-temporal information in the unified feature space, and Method-III focuses on the spatial information to characterize key-frames and objects in the feature space. Both Methods-II and -III produce more representative key-frame sets for object segmentation than Method-I, as demonstrated by the segmentation results.

6. CONCLUSIONS

This paper presents a coherent approach to extract video key-frames for robust object segmentation within a video shot. Key-frame extraction is formulated as a feature selection process that aims at maximizing two divergence-based criteria, i.e., MAIKLD and MMD, in the unified feature space. In the context of object segmentation, the proposed approach explicitly reveals the inherent relationship between key-frames and objects in a video shot. We are developing an analytical approach to further address this issue.

7. REFERENCES

[1] H. Greenspan, J. Goldberger, and A. Mayer, "A probabilistic framework for spatio-temporal video representation and indexing," in *Proc. European Conf. on Computer Vision*, Berlin, Germany, 2002, vol. 4, pp. 461–475.

[2] H. Greenspan, J. Goldberger, and A. Mayer, "Probabilistic space-time video modeling via piecewise GMM," *IEEE Trans. Pattern Analysis and Machine Intelligence*, , no. 3, pp. 384–396, March 2004.

[3] L. Liu and G. Fan, "Combined key-frame extraction and object-based video segmentation," *IEEE Trans. Circuits and System for Video Technology*, 2005, to appear.

[4] N. Vasconcelos, "Feature selection by maximum marginal diversity: optimality and implications for visual recognition," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Madison, Wisconsin, 2003.

[5] S. Kullback, *Information Theory and Statistics*, Dover, New York, 1968.

[6] H. P. Decell and J. A. Quirein, "An iterative approach to the feature selection problem," in *Proc. of Purdue Univ. Conf. on Machine Processing of Remotely Sensed Data*, 1972, vol. 1, pp. 3B1–3B12.

[7] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, pp. 1119–1125, Nov. 1994.

[8] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proc. of IEEE Int Conf on Image Processing*, Chicago, IL, 1998, pp. 866–870.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, vol. 39, pp. 1–38, 1977.

[10] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Annals of Statistics*, vol. 11, no. 2, pp. 417–431, 1983.

[11] R. Linsker, "Self-organization in a perceptual network," *IEEE Computer*, vol. 21, no. 3, pp. 105–117, 1988.