21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Joint Shot Boundary Detection and Key Frame Extraction

Xiao Liu[1], Mingli Song[1], Luming Zhang[1], Senlin Wang[1]
Jiajun Bu[1], Chun Chen[1] and Dacheng Tao[2]
[1] *Zhejiang Provincial Key Laboratory of Service Robot*
College of Computer Science, Zhejiang University
{ender_liux, brooksong, snail_wang, bjj, chenc}@zju.edu.cn
[2] *Center for Quantum Computation and Information Systems, UTS*
dacheng.tao@gmail.com

Abstract

Representing a video by a set of key frames is useful for efficient video browsing and retrieving. But key frame extraction keeps a challenge in the computer vision field. In this paper, we propose a joint framework to integrate both shot boundary detection and key frame extraction, wherein three probabilistic components are taken into account, i.e. the prior of the key frames, the conditional probability of shot boundaries and the conditional probability of each video frame. Thus the key frame extraction is treated as a Maximum A Posteriori which can be solved by adopting alternate strategy. Experimental results show that the proposed method preserves the scene level structure and extracts key frames that are representative and discriminative.

1 Introduction

There are millions of cameras over the world capturing a gigantic amount of video data every day and raises a new challenge: the mass storage and frequent retrieval lead to temp-spatial cost inevitably. Hence it is valuable to allow people to retrieve or gain certain perspectives of a video without watching all the video data.

To maximally transfer the cues from the video into a limited number of key frames, Zhang et al. [5] proposed selecting a key frame if its histogram significantly differs from the previous selected one. This method fails to guarantee the representativeness of the key frames. By representing each frame as a color histogram, Zhuang et al. [6] clustered the frames of a video into several clusters, and further obtained a key frame to describe each cluster. This algorithm totally ignored the temporal information, which is very important for key frame representation. Won et al. [4] detected video shot boundaries using the luminance variance. Their method is based on the difference of modelling errors of an ideally modelled transition. Cernekov et al. [1] firstly detected shots and then extracted key frames using mutual information and the joint entropy. Kelm et al. [2] segmented video into shots by detecting gradual and abrupt cuts, and extracted key frames using visual attention features. Sun et al. [3] extracted key frames at the peaks of the distance curve of color distribution between frames. These methods rely on effective shot boundary detection. Unfortunately, shot boundary detection is data dependent, and it is difficult to obtain accurate detection on different videos. Furthermore, even having gotten semantic shots, the algorithms cannot guarantee each shot involves a unique qualified key frame.

In contrast to the previous algorithms which firstly detect shot boundaries and then extract key frames based on the division, to solve or at least reduce the aforementioned problems, we propose a joint framework to integrate both shot boundary detection and key frame extraction by a probabilistic model. The proposed algorithm is designed to divide a video into fixed number of shots and to select a key frame for each shot such that the selected key frames are best matching to the original video. This formulation enables the shot boundary detection and key frame extraction benefit from each other. And the key frame extraction is treated as a Maximum A Posterior problem which can be solved by adopting alternate strategy.

2 A Probabilistic Model for Representing Video by Key Frames

For frame-based video summarization and retrieval, shot boundary detection is usually taken as the first

978-4-9906441-1-6 ©2012 IAPR

step before key frame extraction. And these two processes are carried out independently. Different from such conventional approaches, we unify them by using a probabilistic model. As illustrated in Figure 1, both shots and key frames are taken into account to represent a video. And by the chain rule of probability, a joint framework can be defined as $P(S, K, V) = P(K)P(S|K)P(V|S, K)$.

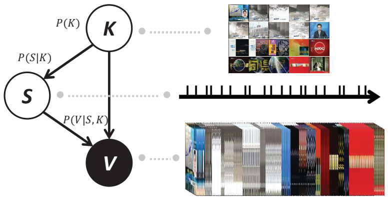


Figure 1. The proposed joint framework.

Given an observed video $V = \{v_i, i = 1 \dots N\}$ with N frames, the goal is to detect M shot boundaries $S = \{s_i, i = 0 \dots M\}$ and extract key frames $K = \{k_i, i = 1 \dots M\}$ from the shots. s_i is the i^{th} shot boundary and the frames with index between s_{i-1} and s_i belong to the i^{th} shot. s_0 and s_M are always set to be 1 and $N + 1$ respectively. The joint framework can be formulated by the following MAP framework:

$$(S^*, K^*) = \max_{(S,K)} P(S, K|V) \tag{1}$$

where $P(S, K|V) = P(K)P(S|K)P(V|S, K)$. $P(K)$ is the key frame prior, $P(S|K)$ is the conditional probability of shot boundaries, and $P(V|S, K)$ is the conditional probability of a video. Since the choices of K and S are coupled, it is intractable to compute them in general. We can adopt an alternate optimization strategy, and solve the problem by iterating the following two updating steps:

$$S^{(t+1)} = \operatorname*{argmax}_{S} P(S|K^{(t)})P(V|K^{(t)}, S) \tag{2}$$

$$K^{(t+1)} = \operatorname*{argmax}_{K} P(K)P(S^{(t)}|K)P(V|K, S^{(t)}) \tag{3}$$

Eq.2 and Eq.3 respectively correspond to shot boundaries detection and key frames extraction. We firstly describe their common factor $P(V|S, K)$, and then give a detailed explanation of these two equations in Section 2.1 and 2.2 respectively.

$P(V|S, K)$ is the conditional probability of video from its key frames and shot boundaries. We assume that a frame in the original video, based on its shot, is independently matched to a key frame:

$$P(V|S, K) = \prod_{i=1}^{M} \prod_{j=S_{i-1}}^{S_i-1} P(v_j|k_i) \tag{4}$$

The conditional probability of a single frame is defined as:

$$P(v_j|k_i) = \exp(-\lambda_g d(v_j, k_i)) \tag{5}$$

where $d(v_j, k_i)$ is the Chi-square distance between the color and texture histograms of two frames. The similar the two frames, the smaller the distance, the greater the generation probability. λ_g is a scale factor. We can rewrite Eq.4 into energy form:

$$P(V|S, K) = \exp(-E_g(V|S, K)) \tag{6}$$

where

$$E_g(V|S, K) = \sum_{i=1}^{M} E_{g,i^-}(V|k_i) + E_{g,i^+}(V|k_i) \tag{7}$$

$$E_{g,i^-}(V|k_i) = \lambda_g \sum_{j=s_{i-1}}^{L(k_i)-1} d(v_j, k_i) \tag{8}$$

and

$$E_{g,i^+}(V|k_i) = \lambda_g \sum_{j=L(k_i)}^{s_i-1} d(v_j, k_i) \tag{9}$$

where $L(k_i)$ indicates the location of the i^{th} key frame in the video, thus: $v_{L(k_i)} = k_i$.

2.1 Shot Boundary Detection

Shot boundary detection is based on Eq.2, which has two factors: $P(S|K)$ and $P(V|S, K)$. The later one, as mentioned before, can be written in an energy form. We here analyze the conditional probability $P(S|K)$.

As defined before, the i^{th} shot boundary is located between k_i and k_{i+1} . Given k_i and k_{i+1}, the choice of s_i is assumed conditional independent:

$$P(S|K) = \prod_{i=1}^{M-1} P(s_i|k_i, k_{i+1}) \tag{10}$$

Furthermore, since the prior of the conditional probability $P(s_i|k_i, k_{i+1})$ is unknown, we simply assume that given k_{i-1} and k_i, s_i is uniformly chosen between the locations of two key frames:

$$P(s_i|k_i, k_{i+1}) = \begin{cases} \frac{1}{L(k_{i+1})-L(k_i)} & L(k_i) < s_i \leq L(k_{i+1}) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

We can rewrite Eq.10 into energy form:

$$P(S|K) = \exp(-E_s(S|K)) \tag{12}$$

where $E_s(S|K) = \sum_{i=1}^{M-1} E_{s,i}(s_i|k_i, k_{i+1})$ and

$$E_{s,i}(s_i|k_i, k_{i+1}) = \tag{13}$$
$$\begin{cases} \frac{1}{\log(L(k_{i+1})-L(k_i))} & L(k_i) < s_i \leq L(k_{i+1}) \\ \infty & \text{otherwise} \end{cases}$$

When K is fixed to be $K^{(t)}$, the minimization of Eq.2 is equal to minimize the energy form:

$$S^{(t+1)} = \underset{S}{\operatorname{argmin}} \Big(E_g(V|K^{(t)}, S) + E_s(S|K^{(t)}) \Big) \tag{14}$$

and

$$\forall i \in \{1 \dots M-1\}, \quad s_i^{(t+1)} = \tag{15}$$
$$\underset{s_i}{\operatorname{argmin}} \Big(E_{g,i^-}(V|k_{i+1}^{(t)}) +$$
$$(E_{g,i^+}(V|k_i^{(t)}) + E_{s,i}(s_i|k_i^{(t)}) \Big)$$

which can be solved through a linear search.

2.2 Key Frame Extraction

Key frame extraction is based on Eq.3, which has three factors: $P(K)$, $P(S|K)$ and $P(V|K,S)$. Since we have no prior knowledge about the key frames, we can firstly apply Bayesian theorem:

$$P(K)P(S|K) = P(K|S)P(S) \tag{16}$$

When S is fixed, $P(S^{(t)})$ is a constant, and we only need to analyze $P(K|S)$. Generally, successive similar key frames are regarded redundant, so we want a key frame is discriminative enough when compared with its adjacent next and previous frames. Then $P(K|S)$ can be formulated as the following Markov Random Field:

$$P(K|S) = \frac{1}{Z_b} \prod_{i=1}^{M} \delta_i \cdot e^{d(k_{i-1},k_i)+d(k_{i+1},k_i)} \tag{17}$$

$$\delta_i = \begin{cases} \frac{1}{s_i - s_{i-1}} & s_{i-1} \leq L(k_i) < s_i \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

Here δ_i is constructed to constrain that the i^{th} key frame is selected from the i^{th} shot. We can rewrite Eq.17 into energy form:

$$P(K|S) = \exp(-E_u(K|S) - E_d(K) - \log Z_b) \tag{19}$$

where

$$E_u(K|S) = \sum_{i=1}^{M} E_{u,i}(k_i|s_{i-1}, s_i) \tag{20}$$

$$E_{u,i}(k_i|s_{i-1}, s_i) = \tag{21}$$
$$\begin{cases} \log(s_i - s_{i-1}) & s_{i-1} \leq L(k_i) < s_i \\ 0 & \text{otherwise} \end{cases}$$

and

$$E_d(K) = \sum_{i=1}^{M} E_{d,i}(k_i|k_{i-1}, k_{i+1}) \tag{22}$$

$$E_{d,i}(S_i|S_{i-1}, S_{i+1}) = -d(k_{i-1}, k_i) - d(k_{i+1}, k_i) \tag{23}$$

To optimize K is equal to minimize the following energy function:

$$K^{(t+1)} = \underset{K}{\operatorname{argmin}} \Big(E_u(K|S^{(t)}) + \tag{24}$$
$$E_d(K) + E_g(V|K, S^{(t)}) \Big)$$

Algorithm 1 Gibbs Sampling for Key Frame Extraction

```
Input: $V$, $S^{(t)}$.
Output: $K$.
Initialize the key frames:
    $\bar{k_i} = \underset{k_i}{\operatorname{argmin}} E_{g,i^-}(V|k_i) + E_{g,i^+}(V|k_i) + E_{u,i}(k_i|s_{i-1}^{(t)}, s_i^{(t)})$
repeat
    for each frame in each shot do
        Calculate its posteriori as a key frame:
            $P(k_i = v_j|\bar{k_{i-1}}, \bar{k_{i+1}}, S^{(t)}, V) = \frac{1}{Z_i} \exp\Big(-E_{g,i^-}(V|k_i) - E_{g,i^+}(V|k_i) - E_{u,i}(k_i|s_{i-1}^{(t)}, s_i^{(t)}) - E_{d,i}(k_i, \bar{k_{i-1}}, \bar{k_{i+1}})\Big)$
        where $\frac{1}{Z_i}$ is the normalized factor.
    end for
    Discretely Sample a frame as the key frame for each shot.
    Update $\bar{K}$ to the new selected ones
until no $\bar{k_i}$ changes or the algorithm reaches enough iterations
Return $K^{(t+1)} = \bar{K}$
```

We use a Gibbs Sampling algorithm which calculates a posteriori for each candidate position and samples the key frames following their posteriori probability distributions.

sequence	F.N.	S.N.	AC.	AC. of [2]
news report	2500	25	98.3	90.5
traffic surveillance	3000	30	80.3	65.5
sports	1500	15	82.3	72.5
movie	10000	100	62.1	42.9

Table 1. The information and the quantitative comparisons on four sequences.

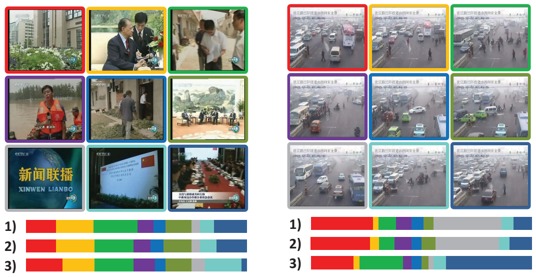
3 Results and Discussions

We tested the proposed algorithm on 4 different video sequences, including a news report sequence, a traffic surveillance sequence, a sports sequence and a movie sequence. The number of key frames/shots is set to be 1% of the total frame number. We can then get the segmentation accuracy of the algorithm through comparing with the human annotation ground truth. Frame numbers (F.N.), shot numbers (S.N.), the accuracy (Ac.) of the proposed method and the accuracy of a independent shot boundaries detection algorithm [2] are shown in Table 1.

Part of the key frames of the TV news report sequence are shown in Figure 3(a). Figure 3(b) shows the key frames of the traffic surveillance sequence. Although the shot boundaries of the traffic surveillance scene are not apparent and it is harder to use key frames to represent this video sequence, our algorithm can extract representative and discriminative key frames. The first key frame corresponds to parallel vehicle flows, while the second key frame is associated to the phase when the right traffic queue ends. The third to the fourth key frames represent that pedestrians walk across the right side and wait in the middle of the street. The fifth key frame shows that the left traffic queue ends and the pedestrians walk across the left side of the street. The sixth key frame represents that some vehicles in the right side turn to the left after the pedestrians. A new cycle begins from the seventh key frame. Compared to the previous independent shot boundaries detection method, e.g. [2], the proposed algorithm is more consistent with the results provided by a human annotator.

4 Conclusions

Aiming at representing a video by its key frame set, we propose a novel probabilistic framework to unify the detection of shot boundaries and the extraction of key frames. The proposed algorithm can automatically divide a video into semantic shots and extract a key frame for each of the shot. Our method outperforms the previous ones by coupling the processes of shot boundaries detection and key frames extraction.



(a) The news report sequence.

(b) The traffic surveillance sequence.

Figure 2. Comparison results: (top) example key frames and (1) shot boundaries detected by the proposed method, (2) human annotation ground truth and (3) independent shot boundaries detection [2].

ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China(61170142), National Key Technology R&D Program (2011BAG05B04), the Zhejiang Province Key S&T Innovation Group Project (2009R50009), and the Fundamental Research Funds for the Central Universities(2012FZA5017)

References

[1] Z. Cernekov, I. Pitas, and C. Nikou. Information theory-based shot cut/fade detection and video summarization. *IEEE Transactions on Circuits and Systems for VIdeo Technology*, (1), 2006.

[2] P. Kelm, S. Schmiedeke, and T. Sikora. Feature-based video key frame extraction for low quality video sequences. *10th Workshop on WIAMIS*, pages 25–28, May 2009.

[3] Z. Sun, K. Jia, and H. Chen. Video key frame extraction based on spatial-temporal color distribution. *Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 196–199, 2008.

[4] J. U. Won, Y. S. Chung, I. S. Kim, J. G. Choi, and K. H. Park. Correlation based video-dissolve detection. *Information Technology: Research and Education*, pages 104–107, 2003.

[5] H. Zhang, J. Wu, D. Zhong, and S. Smoliar. An integrated system for content-based video retrieval and browsing. *Pattern Recog.*, 30:643–658, 1997.

[6] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra. Adaptive key frame extraction using unsupervised clustering. *Proc. ICIP*, 1:866–870, 1998.