INFORMATION-THEORETIC CONTENT SELECTION FOR AUTOMATED HOME VIDEO EDITING

Patricia P. Wang, Tao Wang, Jianguo Li, Yimin Zhang

Intel China Research Center, Beijing, 100080, China

ABSTRACT

In automated home video editing, selecting out the most informative contents from the redundant footage is challenging. This paper proposes an information-theoretic approach to content selection by exploring the dependence relations between who (characters) and where (scenes) in the video. First the footage is segmented into basic units about the same characters at the same scene. To compactly represent the dependence relations between scenes and characters, contingency table is used to model their co-occurrence statistics. Suppose the contents about which characters at which scene are dominating by two random variables, an optimal selection criterion is proposed based on joint entropy. To improve the computation efficiency, a pruned N-Best heuristic algorithm is presented to search the most informative video units. Experimental results demonstrated the proposed approach is flexible and effective for automated content selection.

Index Terms— Automated home video editing, content selection, co-occurrence statistics, information-theoretic

1. INTRODUCTION

Personal video authoring has been a major research issue for multimedia content indexing and management. Existing automated home video editing systems are generally composed of three main steps: analysis, selection, and composition [1]. Selection, whose target is to pick out the most informative and important contents from the long and redundant raw footage, is the primary and challenging step to release end-users from the manual-labored editing work.

For professional edited video, numerous attempts have been made over the years to find the most representative contents from a large collection. For instance, applying Singular Value Decomposition to refine the feature space and to create news report summary [2]; using a temporal graph to model the dynamic evolution of video stream and to produce carton highlights [3]; defining a utility model about audio-visual complexity and grammar for film skimming [4]. For unprofessional home video, although Lienhart has pointed out in earlier years that the good quality rules include balanced coverage, shortened shots, focused selection, and variable editing patterns [5], further investigations were deployed more recently. For instance, Hua et al. proposed a non-linear programming algorithm to formulate the problem of attentional highlight selection [1], Zhao et al. combined audio and visual cues together to pick out event about laughter, applause, and scream as the abstraction results [6].

Selecting the most important and informative content from home video is much harder than from professional video like sports, news or movie, as home video suffers from the loose structure and absence of storyline. Basically, the selection result is determined by three key issues: i) selected unit, shots in home video are usually too long with lots of redundancy, while sub-shots still lack of clear and agreed definition [1]; ii) optimized criteria, what content is more important than others greatly depends on the subjective judgment, while picking out the most informative content still lack of theoretical solution more than assessment rules [5]; and iii) implemented strategy, without pair-wise comparison and iterative search, it is hard to tell how long and which content should be included even for end-users [3].

With investigation, we found that for issue i) end-users generally are interested in whom and where involved in the video, therefore it is acceptable to treat the contents about the same characters at the same scene as the selected units without loss of generality; ii) most personal video capturer would like the contents about every character at every scene be completely and uniformly included, leaving special request behind, thus it is reasonable to measure the redundancy of information as the optimized criterion; iii) finding the most informative content through pair-wise comparison and iterative search is a combinatorial computational problem, accordingly it is appropriate to design a heuristic search algorithm to approximate the user selection behavior.

In this paper, we propose a novel approach to content selection for automated home video editing. The approach proceeds by segmenting the raw footage into scenes taken at the same place, where each scene is further segmented into subscenes capturing the same characters. Based on these basic units, we represent the dependence relations between characters and scenes by modeling their co-occurrence statistics with a contingency table. Then by treating whom at where in the video are dominated by two random variables, we apply joint entropy to measure the redundancy of information conditional on a set of video units. Finally, to improve the computation efficiency, we present a pruned N-Best heuristic search algorithm to optimize the selection criterion. Experimental results demonstrated that the proposed approach is able to adaptively select the most informative content while guarantee the perceived quality of edited video.

2. CO-OCCURRENCE STATISTICS MODELING OF SCENES AND CHARACTERS

The elementary step to content selection is to generate the basic units that have consistent semantic and appropriate length. As who and where are the two main subjects concerned by personal video capturers, it is reasonable to segment the raw footage in terms of scenes (where) and characters (who), based on which we get the basic units about the same characters at the same scene. Considering there are many-to-many correlations between which character and which scene, it is valuable to represent the content by grasping their co-occurrence statistics. In this section, we propose a compact representation of home video by modeling the co-occurrence statistics between scenes and characters.

2.1. Scene and Character-based Video Segmentation

In home video, we define a scene as the situation where a particular set of activities happened, and refer to a character as the person who is involved in a particular set of scenes. It is applicable to define the contents about the same characters at the same scene as the basic units to be selected for automated editing. To get such units, we need to perform the scene detection and face recognition. It is aware that the same scenes usually have similar background and are temporally adjacent, and the characters are always focused on several family members and captured by frontal close-up. Therefore in implementation, we extracted the global color histogram for each frame and measure their spatio-temporal coherence to segment home video into various scenes [3]. Besides, we extracted the SIFT and Gabor features to track the head movement, and identify the near frontal face to classify the head tracking results into various characters [7]. The scene and character-based video segmentation results are further processed and combined to get the basic units describing the same characters at the same scene. Such video unit has consistent semantic and appropriate length readily to be used for content selection.

At the left of Fig.1, we show an example of the basic units describing the same characters at the same scene. Here, two scenes and three characters have been detected. According to the appearance of characters within each scene, we then generate four video units: $\{u_1, u_2, u_3, u_4\}$. Let $\mathbf{S} = \{\mathbf{s}_k, 1 \leq k \leq K\}$ and $\mathbf{C} = \{\mathbf{c}_l, 1 \leq l \leq L\}$ denote the scene set, and character set respectively. Then, let $\mathbf{U} = \{u_n, 1 \leq n \leq N_U\}$ denotes the resulting video unit set. For each unit u_n, we use a triple (d_n, s_n, C_n) to describe its duration, scene identity, and character identities, where $s_n \in \mathbf{S}$, $C_n \in 2^{\mathbf{C}}$. Note there may be several characters appear within one unit, thus C_n describes a set of character identities.

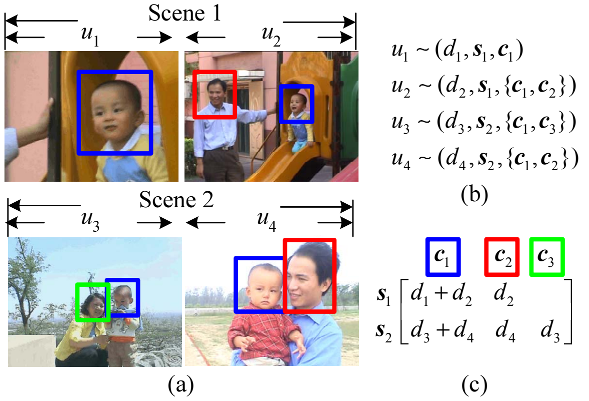
Scene 1
u_1
u_2
Scene 2
u_3
u_4
(a)
$u_1 \sim (d_1, \mathbf{s}_1, \mathbf{c}_1)$
$u_2 \sim (d_2, \mathbf{s}_1, \{\mathbf{c}_1, \mathbf{c}_2\})$
$u_3 \sim (d_3, \mathbf{s}_2, \{\mathbf{c}_1, \mathbf{c}_3\})$
$u_4 \sim (d_4, \mathbf{s}_2, \{\mathbf{c}_1, \mathbf{c}_2\})$
(b)
$\mathbf{c}_1$ $\mathbf{c}_2$ $\mathbf{c}_3$
$\mathbf{s}_1$ $d_1 + d_2$ d_2
$\mathbf{s}_2$ $d_3 + d_4$ d_4 d_3
(c)

Fig. 1. An example of (a) basic units describing the same characters at the same scene, (b) description for each unit, and (c) contingency table construction.

2.2. Co-occurrence Statistics Modeling

Usually, a character may appear in different scenes, and a scene may include different characters. There are many-to-many associations between scenes and characters. Let the scene and the character be two observable attributes of video content, modeling the co-occurrence statistics between them is able to represent the content from a compact point of view. As the two observable attributes takes value from discrete sets **S** and **C**, it is appropriate to use a two-dimensional contingency table to model their co-occurrence statistics. Considering the duration about which character at which scene is an important factor to represent the home video content, it is necessary to count the frame-level attributes when constructing the contingency table.

Based on the detected number of scenes K and the number of characters L, we construct a K-by-L table $M = \{m_{kl}\}$. Given the unit set $\mathbf{U}$ and each unit u_n, we calculate the value of each entry as follows: $m_{kl} = \sum_{n|s_n = \mathbf{s}_k, \mathbf{c}_l \in C_n} d_n$. An example is illustrated at the right of Fig.1, where $K = 2$, $L = 3$. After going through all the units in $\mathbf{U}$, the contingency table represent the video content about various scenes and various characters from a statistical perspective. For instance, each entry m_{kl} describes the co-occurrence duration about the character $\mathbf{c}_l$ at the scene $\mathbf{s}_k$, the sum of l^{th} column describes the occurrence duration about the character $\mathbf{c}_l$.

3. INFORMATION-THEORETIC CRITERION FOR AUTOMATED CONTENT SELECTION

After the raw footage is segmented into a set of units $\mathbf{U}$, the next crucial step for automated content selection is to define a reasonable criterion and to find an optimal subset from **U**. Based on the compact representation *M* of the co-occurrence statistics between scenes and characters, in this section we first formulate the selection problem from an information theoretic perspective, then we propose a pruned heuristic algorithm to improve the searching efficiency.

3.1. Joint Entropy based Redundancy Measure

After the home video is described by attributes of the scene and the character, we are able to consider the video content as being dominated by two random variables *X* and *Y*. Let $P(X, Y)$ denotes their joint distribution, as *X* and *Y* are discrete variables which take value from **S** and **C**, we could use their co-occurrence statistics to approximate $P(X, Y)$. Recall the constructed contingency table *M* whose element m_{kl} describes the co-occurrence duration about the character $\mathbf{c}_l$ at the scene $\mathbf{s}_k$, we may calculate $P(X, Y)$ as follows:

$$P(X = \mathbf{s}_k, Y = \mathbf{c}_l) = \frac{m_{kl}}{\sum_k \sum_l m_{kl}}, \tag{1}$$

where for simplicity let p_{kl} denotes $P(X = \mathbf{s}_k, Y = \mathbf{c}_l)$.

The target of content selection is to measure the amount of information and redundancy conveying from set **U**. As the joint entropy [8] measures the average information associated with multiple random variables, it is reasonable to measure the amount of information associated with a scene and a character:

$$H(X, Y) = -\sum_k \sum_l p_{kl} \times \log p_{kl}. \tag{2}$$

In our case, the optimal criterion means that different scenes and different characters have been completely and uniformly selected, so that maximum amount of information is conveyed as well as the redundancy is eliminated. Generally, the maximum joint entropy is achieved when the joint probability p_{kl} is uniformly distributed. In this view, the optimal criterion for content selection is to find a subset U_i of **U** whose joint probability distributed as uniformly as possible, i.e.,

$$\tilde{U} = \arg\max_{U_i}(H(X, Y|U_i)), U_i \subset \mathbf{U}. \tag{3}$$

3.2. Pruned N-Best Heuristic Search Algorithm

Giving the optimal criterion in (3), the content selection could be treated as a search problem. As we know, brute-force algorithm could determine the global optimal, but it is NP-complete and a combinatorial computational problem. To improve the search efficiency, we investigated the user behaviors of manual selection. It is observed that end-users generally go through all the video content and exclude the most redundant one, then examine the remainders and exclude the secondary redundant. Such procedure is repeated until the remainders convey the most amount of information and eliminate most redundancy. Based on this investigation, we present an iterative search algorithm which examines all the possible subsets U_i at each step, and remains the N-Best subsets Φ for the next step examination. As the subsets excluded from the previous step will not be examined at the next step, it is a pruned N-Best heuristic search algorithm.

```
Define: $U^{n-} = \mathbf{U} - \{u_n\}, 1 \leq n \leq N_U$.
Initialize: $i = 0, \Phi_0 = \mathbf{U}$.

  1. $i = i + 1, \Phi_* = \oslash$;
  2. for each $U_{i-1}^{n-} \in \Phi_{i-1}, 1 \leq n \leq N_{U_{i-1}^{n-}}$
  3.         $\Phi_* = \Phi_* \cup \arg\max_{U_{i-1}^{n-}}^{(N)} \{H(X, Y|U_{i-1}^{n-})\}$;
  4. end;
  5. $\Phi_i = \arg\max_{U_*}^{(N)} \{H(X, Y|U_*)\}$, where $U_* \in \Phi_*$;
  6. if $\exists U_i \in \Phi_i$, and $U_i$ satisfy the selection ratio, stop;
  7. else go to step 1.
```

Fig. 2. Pseudo-code of the proposed pruned N-Best heuristic search algorithm.

The pseudo-code of the proposed algorithm is presented above. At step 3, $\arg\max^{(N)}$ means finding the arguments that lead to the N-best maximum value. After i^{th} iteration, Φ_i will output *N* subsets with N-Best maximum joint entropy. When calculating the joint entropy at step 3, the joint distribution $P(X, Y|U_{i-1}^{n-})$ is easy to update from $P(X, Y|U_{i-})$ by taking into account the duration d_n of unit u_n. In our experiments, we validate that $N = 1$ could meet the practical needs when balancing the algorithm robustness and search efficiency.

4. EXPERIMENTS AND DISCUSSIONS

To test the performance of the proposed approach, about 4-hour home video, in MPEG-2 format with 768 × 576 size (PAL), have been used. The evaluation is carried out for segmentation accuracy, search efficiency, and subjective assessment. Normalized mutual information [8] ($\bar{I}(X, Y) \triangleq \frac{I(X,Y)}{max(H(X),H(Y))} \in [0, 1]$) is used as the metric for segmentation accuracy. Table 1 lists the segmentation results of scene (**S**), character (**C**) and basic unit (**U**), the number of ground truth ($\#g$) and number of detected segments ($\#d$). For scene segmentation, the general performance is good, as the closer $\bar{I}$ is to 1, the segmented result is more consistent with ground truth. For character deetction, the problem of over-classification ($\#d > \#g$) is mainly because of the large variations in face pose. This will lead to various poses of one character be uniformly and completely selected.

	V_1			V_2			V_3		
	S	**C**	**U**	**S**	**C**	**U**	**S**	**C**	**U**
$\#g$	16	8	267	26	3	172	11	5	123
$\#d$	15	27	289	28	9	196	14	13	141
$\bar{I}$	0.82	0.59	0.68	0.86	0.56	0.63	0.89	0.54	0.71

Table 1. Performance of scene (**S**), character (**C**), and basic units (**U**) detection.

To demonstrate the robustness of the proposed N-Best search algorithm, we compare the change of maximum joint entropy given various N. Figure 3 shows an example for $N = 1$ and $N = 2$ on the 267 video units of V_1, where X-axis denotes the number of units selected in the subset. At the beginning, joint entropy is 5.38 and it increases when less units are selected. After 265 iterations, the peak value 8.39 appears, which indicates that 12 video units are enough to capture the dependence between characters and scenes existing in the original 267 units. We see that for $N = 1$ and $N = 2$, the two curves of maximum joint entropy are very close to each other. It indicates that $N = 1$ is able to provide the optimal selection results as well as the computation efficiency is guaranteed.

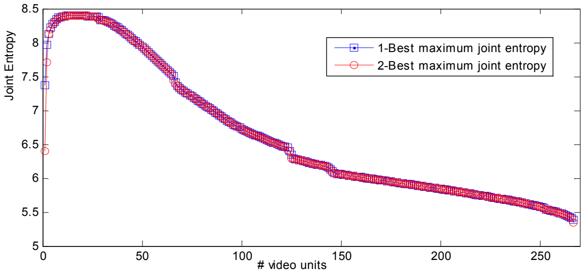

Fig. 3. Illustration of the 1-Best and 2-Best maximum joint entropy on 267 units of video V_1.

Finally, we carried out the subjective evaluation. The proposed information-theoretic approach (**I.**) is compared with the manual selection (**M.**) and random selection (R.), and Table 2 lists the comparative results. Here the selection ratio is set as 50%, 25%, and 10% respectively. We have invited five graduates to assess a score ranging from $[0, 100]$ for each selection result. The satisfaction score is then calculated by averaging five graduates' assessment results. We see that when the selection ratio is lower, the performance of our approach is closer to that of manual selection. When the selection ratio is 10%, our approach intends to select the contents about multiple characters within a scene, which is consistent with the user's behavior in manual selection, i.e., trying to include more characters and more scenes in limited duration.

	V_1			V_2			V_3		
	R.	**M.**	**I.**	**R.**	**M.**	**I.**	**R.**	**M.**	**I.**
50%	66.7	94.0	88.1	63.4	92.8	79.3	53.9	94.9	89.6
25%	60.4	90.8	75.7	65.3	96.9	82.5	58.3	95.8	86.7
10%	56.2	92.8	83.4	63.7	94.1	86.0	58.3	90.0	86.3

Table 2. Subjective evaluation results for random, manual, and the proposed information-theoretic selection approaches when selection ratio is 50%, 25%, and 10%.

Compared the results of manual selection with our approach, we found that low-quality artifacts like blur and shaking badly influence user's selection. In future work, either enhancement or low-quality removal will be included. Besides, end-users pay much attention to what has been done by the characters. Therefore, investigation of character's movement would be significant to identify the interesting content.

5. CONCLUSIONS

This paper proposed an information-theoretic approach to content selection for automated home video editing. In this approach, the raw footage is temporally segmented into video units in terms of characters and scenes which are the two main components cared for by general end-users. To enable the video content about every character at every scene be uniformly and completely included in the edited video, joint entropy is employed to measure the amount of information associated between scenes and characters. To get the optimal selection result, we proposed a robust and efficient searching algorithm. Experimental results have demonstrated the performance of the proposed approach in terms of segmentation accuracy, search efficiency, and subjective assessment.

6. REFERENCES

[1] X.-S. Hua, L. Lu, and H.-J. Zhang, "Ave - automated home video editing," in *ACM Multimedia*, 2003, pp. 490–497.

[2] Y. Gong and X. Liu, "Video summarization using singular value decomposition," in *CVPR*, 2000, vol. 2, pp. 174–180.

[3] Z. Rasheed and M. Shah, "Detection and representation of scenes in videos," in *IEEE Transactions on Multimedia*, 2005, vol. 7, pp. 1097–1105.

[4] H. Sundaram, L. Xie, and S.-F. Chang, "A utility framework for the automatic generation of audio-visual skims," in *ACM Multimedia*, 2002, pp. 189–198.

[5] R. Lienhart, "Abstracting home video automatically," in *ACM Multimedia*, 1999, pp. 37–40.

[6] M. Zhao, J. Bu, and C. Chen, "Audio and video combined for home video abstraction," in *ICASSP*, 2003, vol. 5, pp. 620–623.

[7] W. Fan and T. Wang *et al.*, "Semi-supervised cast indexing for feature-length films," in *MMM*, 2007.

[8] T.M. Cover and J.A. Thomas, "Elements of information theory," 1991, Wiley, New York.