Video Summarization via Transferrable Structured Learning

Liangda Li
College of Computing
Georgia Institute of Technology
Atlanta, GA 30032
ldli@cc.gatech.edu

Ke Zhou
College of Computing
Georgia Institute of Technology
Atlanta, GA 30032
kzhou@cc.gatech.edu

Gui-Rong Xue*
Alibaba Group R&D
No. 391, Wen'er Road,
Hangzhou, China 310099
grxue@aliyun-inc.com

Hongyuan Zha
College of Computing
Georgia Institute of Technology
Atlanta, GA 30032
zha@cc.gatech.edu

Yong Yu
Dept. of Computer Science and Engineering
Shanghai Jiao-Tong University
No. 800, Dongchuan Road,
Shanghai, China 200240
yyu@apex.sjtu.edu.cn

ABSTRACT

It is well-known that textual information such as video transcripts and video reviews can significantly enhance the performance of video summarization algorithms. Unfortunately, many videos on the Web such as those from the popular video sharing site YouTube do not have useful textual information. The goal of this paper is to propose a transfer learning framework for video summarization: in the training process both the video features and textual features are exploited to train a summarization algorithm while for summarizing a new video only its video features are utilized. The basic idea is to explore the transferability between videos and their corresponding textual information. Based on the assumption that video features and textual features are highly correlated with each other, we can transfer textual information into knowledge on summarization using video information only. In particular, we formulate the video summarization problem as that of learning a mapping from a set of shots of a video to a subset of the shots using the general framework of SVM-based structured learning. Textual information is transferred by encoding them into a set of constraints used in the structured learning process which tend to provide a more detailed and accurate characterization of the different subsets of shots. Experimental results show significant performance improvement of our approach and demonstrate the utility of textual information for enhancing video summarization.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing-abstracts methods

General Terms

Algorithm, Experimentation, Performance

Keywords

Video Summarization, Transfer Learning, Structural SVM

1. INTRODUCTION

Video summarization is the process of generating the montage of a given video that indicates its main theme and contents. Nowadays, the fast development of the Web has resulted in the explosive growth of video resources, which makes video summarization a very important technology for efficient access of video contents. Among those large amount of Web video resources, some are accompanied with additional types of information, such as texts, and the exploration of textual information has been shown to benefit the video summarization task [20, 12, 31]. However, there are also significant amount of videos that do not possess those extra information. For them, summarization models relying on both video and text features cannot be applied directly.

Early video summarization approaches [20, 5, 33, 29], depicted in Figure 1(a), usually just make use of video resources alone, and conduct the learning process based on video features. Those approaches assume that video features can well determine whether certain shots should be included in an ideal montage, or whether a subset of shots as a whole constitutes an ideal montage. However, due to the complex structure and representation of video shots, it is generally difficult to design effective features from video that can capture the semantic importance of video shots precisely.

More recent approaches [4, 24, 26], depicted in Figure 1(b), realize the importance of other types of resources, especially textual information. Those approaches achieve significant improvement and demonstrate that text features can give a clear and accurate description of the corresponding shots. However, they require both video and text features

*Gui-Rong Xue is the corresponding author.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0632-4/11/03.

Figure 1: Three different strategies for video summarization

in montage prediction. Considering that current video resources, such as those from YouTube [1], are mostly without adequate textual data, the above approaches can only be applied in some limited scenarios.

To address the above mentioned problems, we propose a transfer learning approach, depicted in Figure 1(c). It makes use of both video and textual data in the training process; and video data alone are required in the testing/summarization process. More specifically, we formulate the summarization task as a subset selection problem. As shown in Figure 2, our basic assumption is that in the training process, each shot is accompanied by certain textual information. We can specify two summarization tasks based on video and text, respectively. Both of them actually aim at learning the same structure indicating which candidate subset should be selected to generate the summary. This makes the knowledge transfer from text to video possible: It is observed that there exists a close relationship between videos and their corresponding text materials. Specifically, those video-based features and textual features are highly correlated, which can be described by a transfer matrix. This means that candidate montages probably differ from each other to a similar degree as the corresponding subsets of text information, usually represented in the form of sentence subsets or word subsets. Therefore, by integrating the textual data into the learning process, we can expect a more powerful summarization model containing the knowledge transferred from textual data.

In this paper, we will focus on dealing with a special type of videos of TV series, they tend to have the following characteristics: 1) TV series are usually accompanied with rich textual information such as transcripts and dramas, which are closely related to the content of the TV series; and 2) For each episode of a TV series, it often begins with a montage,

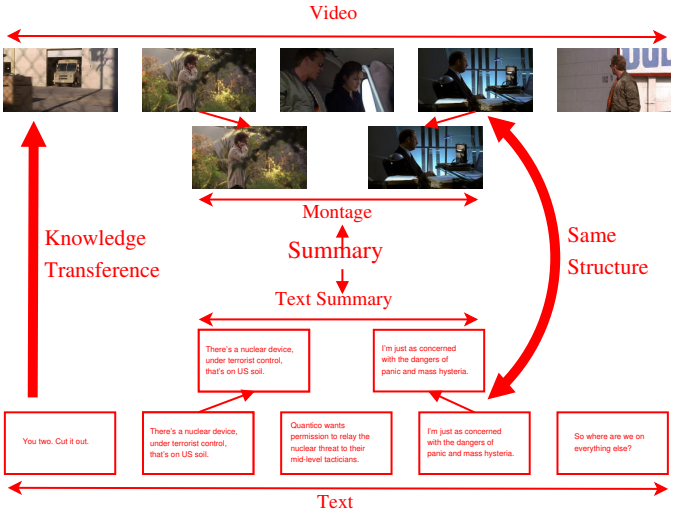

Figure 2: The mechanism of our transfer learning approach

in other words, video summary, of the last episode, thus provides ideal training data for the summarization task.[2] Notice that our assumption relies on the requirement that video and corresponding textual information are closely related. For other types of videos, such as field sports, indoor sports, home movies, whose accompanied textual information is not always so close related to the video contents, different kinds of methods need to be developed.

Our paper aims at making good use of videos' associated textual data — available during the training process — to help the summarization task based on video data only. To

[1]Videos from YouTube are usually accompanied by few content-light comments, brief descriptions of video producers and actors, or short vacuous outlines.

[2]Those popular video data sets, such as TRECVID [2], usually are not accompanied by textual information, and they also lack inner story structures.

achieve our goal, we utilize structural SVM as the backbone of our training method augmented with a set of constraints generated from the textual data. This forms the proposed transferable structured learning framework. In the process of constraint generation, a text summarization model is trained to evaluate the importance of candidate montages in terms of the corresponding textual information. Then through the model, textual data are encoded into constraints which are supposed to give a more detailed description of the difference between candidate montages. Thus, the learning process of the video summarization is guided by the knowledge from the associated text summarization. As a result, the text information is transferred into knowledge on video summarization, enhancing the learning process of video summarization. Cutting plane algorithm is adapted to solve the resulting optimization problem [30]. Finally, in summarizing a new video, the learned summarization model is used for montage prediction using video features only.

A set of experiments have been conducted to demonstrate the effectiveness of the proposed summarization method. Firstly, our approach was shown to give a remarkable improvement compared with several state-of-the-art methods. It also showed that the proposed constraints, which are supposed to transfer the textual data into knowledge on the summarization task based on only video features, contribute to a more effective video summarization model. Finally, we also demonstrated that the effectiveness of the pre-learned text summarization model has a significant impact on the performance of the video summarization task.

We now list the main contributions of our work:

1. We propose the idea of using textual information to enhance video summarization in the context of transfer learning: textual information is only used for training the summarization model, and summarization of new videos relay on video features alone;

2. We propose a transferable structured learning framework, leading to a more effective summarization model, which use constraints to incorporate knowledge transfer from textual data.

The rest of the paper is organized as follows: in Section 2 we present previous works related to video summarization and transfer learning. The video summarization task is formulated as a structured learning problem in Section 3. In Section 4, we discuss in detail how to transfer videos' related textual data into knowledge on how to generate montage just based on video features. Then we address how to solve the resulting optimization problem in Section 5; detailed description of the features we use in the learning process is also included in this section. Section 6 presents the experimental results along with some case studies. Finally, We conclude the paper with some pointers to future research directions in Section 7.

2. RELATED WORKS

2.1 Video Summarization

Summarization [15, 25, 5, 10] has enjoyed considerable development in other areas, such as information retrieval. In particular, video summarization attracts increasingly interests nowadays. It can be generally categorized into keyframe-based summarization and skimming-based summarization.

Keyframe-based summarization has been studied for quite a long time. These approaches usually represent each shot by keyframes and extract a certain number of them or their corresponding shots as the montage of given videos. According to the techniques on keyframes extraction, they can be further classified as follows: sampling-based, shot-based, and graph-based. Sampling-based approaches, such as video magnifier [20] and Minivideo [25], usually extract keyframes from videos in a random or uniform manner. Shot-based approaches firstly segment video into shots and extract one or several keyframes from each of them based on low-level video features, such as color or motion. S. X. Ju et al. [12], once employ motion and gesture as the criterion for keyframe extraction. Advanced human behaviors are also studied to extract more meaningful and representative keyframes. Attention models, employed in [21], prove to be a reasonable criteria for keyframe extraction. Recently, several graph-based approaches [31, 33] have been proposed. These approaches view a given video as a graph with each node representing a standard unit. Then various graph-based algorithms such as clustering, singular value decomposition, or principle component analysis, are applied to extract keyframes.

Skimming-based summarization enjoys a fast development in recent years. These approaches make use of the whole shot's information, employ video features of relatively high level, pay special attention to the montage's length, and mostly perform an extra smoothing step to make the final montage more nature. Amir et al. [5] employ video features based on audio to capture the semantic information of shots. Several rush summarization approaches [29, 8] focus on the fast-moving object and camera events of videos. Moreover, shot-change patterns are also used for dialog events extraction in [16]. Most of those approaches mainly deal with video of special domains, as they can make use of the domains' specific information. For example, Russell [23] employs a presentation structure for video's from weekly forum. Several others approaches also pay special attention to videos of talks [9], news [10] or sports [32].

Those above approaches mostly focus on the generation of various kinds of video features, so as to capture the video's structure or distinguish each shots from each other. In spite of their encouraging achievements, above approaches have to face the great difficulty that video features may fail to capture the shots' characteristics accurately due to the overwhelming difficulties in representing videos effectively. Recent years, other types of resources, especially textual data are employed to help video summarization. Some approaches aim at enhancing the semantic characters of videos by generating concept entity [4] or short synopses [24]. Huang et al. [10] aim at generating semantically meaningful montage by integrating text information. A novel video summarization method proposed by Chen et al. [4] employs text features extracted from speech transcript to generate concept entities. The Informedia Project [24] generates short synopses partly based on pre-extracted keywords. Others [3, 26, 36] exploit transcript information to decide the scenes' boundary. Our approach goes a step further by transferring the text information into the learned model for the summarization task based on video features only. In this way, our approach can be applied to a wide range of videos on the web, even if they are not accompanied by textual data.

2.2 Transfer Learning

Another field closely related to our work is transfer learning. Transfer learning approaches mainly aim at make use of knowledge of source domain to solve problems of target domain. According to the form of transferred knowledge, these approaches can to classified into four categories: instance-based, feature-based, parameter-based and relationship-based. Instance-based transfer learning [7, 11] reuses certain parts of data in source domain to help learning in the target domain in the manner of re-weighting. Feature-based approaches [28] mainly focus on how to transfer the feature representation for different domains. Early approaches usually solve the task by constructing rules of each domain and learning to translate rules of different domains. For example, Rule Transfer, a transfer algorithm proposed in [27], prove to be efficient for transferring knowledge between two kinds of games. In recent approach, knowledge of more kinds of forms is enabled to be transferred. A novel approach, translated learning, [6] employs a dictionary concerning two domains, and learn to transfer knowledge of common features of different domain through a Markov chain. mixture coaching, proposed in [28], learns to transfer feature representation in the opposite direction, from target to source, and makes prediction based on features from both domains. Parameter-based approaches, such as response coaching [28] and informative vector machine [14], assume that there exists some parameters shared by both source and target tasks. Another class of methods, relationship-based transfer learning, assumes that there exists similar relationship among data in source and target domains. For example, Mihalkova et al. [19] employ Markov logic network to capture the possible similar relationship. Yu et al. [34] included latent variables in a structure learning framework, which is applicable to the cases where loss function is independent of latent variables. Our approach belongs to this category, and furthermore, requires no dependence relation between loss function and features belonging to different domains.

3. PROBLEM DEFINITION

We formulate the video summarization problem as follows. Given a video $\mathbf{v} = \{v_1, v_2, \ldots, v_n\} \in \mathcal{V}$, where v_i represents a shot of the video, $\mathcal{V}$ is the space of all videos. The goal of the summarization task is to predict a shot subset $\mathbf{y}$ from the space of all possible shot subsets $\mathcal{Y}$, in other words, $\mathbf{y} \subset \mathbf{v}$. In video summarization, the subset $\mathbf{y}$ is also called a montage of the video $\mathbf{v}$.

In the training process, we assume that certain corresponding textual information $\mathbf{t} = \{t_1, t_2, \ldots, t_m\} \in \mathcal{T}$ of the video $\mathbf{v}$ is also available. Here t_i denotes the corresponding textual data of shot v_i, for example the corresponding transcript of the shot, and $\mathcal{T}$ denotes the space of all the text. Given a video $\mathbf{v}$, a textual summary $\mathbf{l}$ can also be obtained by collecting the corresponding textual data of its montage $\mathbf{y}$. We use $\mathbf{x}$ to denote the video-text pair $(\mathbf{v}, \mathbf{t})$, and $\mathbf{z}$ their corresponding summary $(\mathbf{y}, \mathbf{l})$.

Assume that we are given a set of labeled training data:

$$\{(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) | i = 1, \ldots, n\},$$

$$\mathbf{x}^{(i)} = (\mathbf{v}^{(i)}, \ \mathbf{t}^{(i)}), \ \mathbf{z}^{(i)} = (\mathbf{y}^{(i)}, \mathbf{l}^{(i)}),$$

where $\mathbf{y}^{(i)}$ represents the *ground-truth* video summary of the video $\mathbf{v}^{(i)}$ and $\mathbf{l}^{(i)}$ the corresponding text summary of $\mathbf{t}^{(i)}$. We want to emphasize that in our transfer learning framework, the labeled training data $(\mathbf{x}, \mathbf{z})$ for both video and its corresponding text are used for the training process, while for summarizing a new video, we will generate video summary without using the textual information.

Now we formulate the problem of video summarization as that of learning a discriminant function,

$$\mathcal{F}(\mathbf{v}, \mathbf{y}) = \mathcal{V} \times \mathcal{Y} \mapsto \mathcal{R},$$

which is intended to measure the degree to which the subset $\mathbf{y}$ considered as suitable summary for the video $\mathbf{v}$. Then we generate the summary of a given video $\mathbf{v}$ as the subset $\mathbf{y}^*$, maximizing $\mathcal{F}(\mathbf{v}, \mathbf{y})$ over all possible subsets $\mathbf{y} \in \mathcal{Y}$, i.e.,

$$\mathbf{y}^* = \underset{\mathbf{y} \in \mathcal{Y}}{\operatorname{argmax}} \mathcal{F}(\mathbf{v}, \mathbf{y}). \quad (1)$$

To make the above framework feasible, we represent each pair $(\mathbf{v}, \mathbf{y})$ by a feature vector $\Psi(\mathbf{v}, \mathbf{y})$. For simplicity, the discriminant function $\mathcal{F}(\mathbf{v}, \mathbf{y})$ is assumed to be linear in terms of the feature vector $\Psi(\mathbf{v}, \mathbf{y})$, i.e.,

$$\mathcal{F}(\mathbf{v}, \mathbf{y}) = \mathbf{w}^T \Psi(\mathbf{v}, \mathbf{y}). \quad (2)$$

As will be shown, the weight vector $\mathbf{w}$ is estimated by fitting both the video training data and the text training data. We also employ the following loss function to quantify the penalty of using a predicted summary $\bar{\mathbf{y}}$ as an approximation of the ground-truth summary $\mathbf{y}$:

$$\Delta(\mathbf{y}, \bar{\mathbf{y}}) : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathcal{R}.$$

In our study, a loss function related to F_1 measure is applied:

$$\Delta(\mathbf{y}, \bar{\mathbf{y}}) = 1 - \frac{2pr}{p + r}, \quad p = \frac{< \mathbf{y}, \bar{\mathbf{y}} >}{< \bar{\mathbf{y}}, \bar{\mathbf{y}} >}, \ \ r = \frac{< \mathbf{y}, \bar{\mathbf{y}} >}{< \mathbf{y}, \mathbf{y} >}. \quad (3)$$

Here, given two subsets $\mathbf{a}$ and $\mathbf{b}$, $< \mathbf{a}, \mathbf{b} >$ denotes the number of common elements they share.

4. TRANSFER LEARNING FROM TEXT TO VIDEO

Our general idea is to extend structural SVM into the transfer learning scenario through introducing an additional set of constraints; and these constraints are meant to encode and transfer the knowledge learnt from an auxiliary and related task. First, we describe briefly the structural SVM method proposed in [35, 15] to train a robust model for a summarization task. Given a training set $\{(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) \mid i = 1, \cdots, n\}$, structural SVM is employed to learn a weight vector $\mathbf{w}$ for the discriminant function $\mathcal{F}(\mathbf{v}, \mathbf{y}) = \mathbf{w}^T \Psi(\mathbf{v}, \mathbf{y})$ through the following quadratic programming problem:

Optimization Problem 1. (Structural SVM)

$$\min_{\mathbf{w}, \xi \geq 0} \frac{1}{2} \|\mathbf{w}\|^2 + \frac{c}{n} \sum_{i=1}^{n} \xi_i, \quad (4)$$

subjected to:

$$\forall i, \forall \mathbf{y} \in \mathcal{Y} \setminus \mathbf{y}^{(i)}, \ \xi_i \geq 0,$$
$$\mathbf{w}^T \Psi(\mathbf{v}^{(i)}, \mathbf{y}^{(i)}) \geq \mathbf{w}^T \Psi(\mathbf{v}^{(i)}, \mathbf{y}) + \Delta(\mathbf{y}^{(i)}, \mathbf{y}) - \xi_i.$$

In Equation (4), the parameter c controls the tradeoff between the model complexity $\frac{1}{2}\|\mathbf{w}\|^2$ and $\sum_{i=1}^{n} \xi_i$, the sum of the slack variables ξ_i. The constraints for the optimization problem enforce the requirement that the ground-truth summary $\mathbf{y}^{(i)}$ should have a higher function value than other alternatives $\mathbf{y} \in \mathcal{Y}$, and $\mathbf{y} \neq \mathbf{y}^{(i)}$.

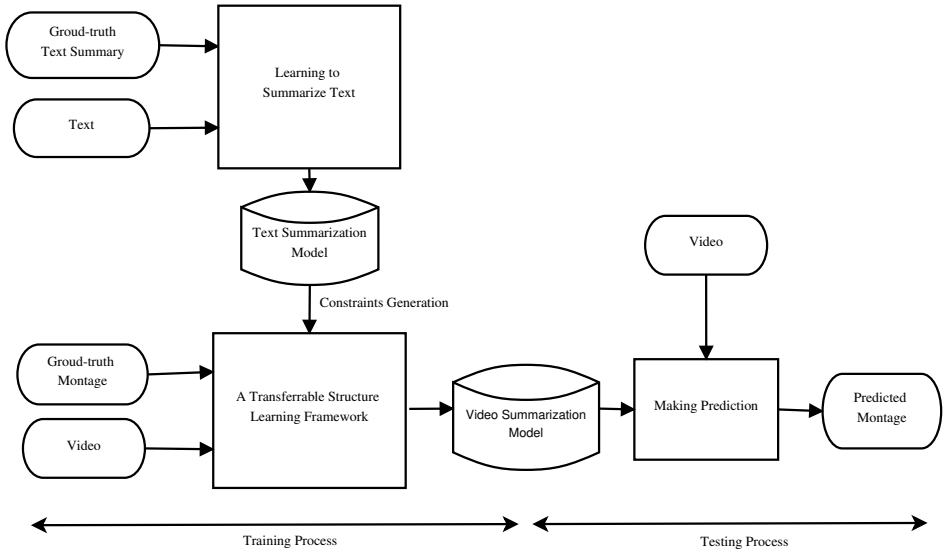

Figure 3: The flowsheet of our video summarization approach

4.1 Text Based Constraints for Video Summarization

Generally, text-based summarization is a much easier task than video-based summarization. Thus, the characteristics of $\mathbf{x} = (\mathbf{v}, \mathbf{t})$ can be better captured by the features generated based on text information. On the other hand, it is quite difficult to obtain effective representation for video shots. Let us consider the constraints in the above-mentioned optimization problem:

$$\mathbf{w}^T \Psi(\mathbf{v}^{(i)}, \mathbf{y}^{(i)}) \geq \mathbf{w}^T \Psi(\mathbf{v}^{(i)}, \mathbf{y}) + \Delta(\mathbf{y}^{(i)}, \mathbf{y}) - \xi_i \qquad (5)$$

These constraints employ $\Delta(\mathbf{y}^{(i)}, \mathbf{y})$ to measure the difference between candidate montages. As mentioned in the previous section, this only describes the difference between two subsets in terms of the F_1 measure. Therefore, it is desirable to more accurately capture the more subtle difference between the ideal subset and an alternative in a quantitative way. The corresponding text materials for video shots, which are known to be closely correlated with the video content, can provide a much detailed and accurate description for the difference between candidate montages as we will show next.

Our basic idea is to use the videos' related textual data to form an additional set of constraints, which is supposed to help training a better video summarization model. Recall that we have the set of training examples,

$$\{(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) | i = 1, \ldots, n\},$$

$$\mathbf{x}^{(i)} = (\mathbf{v}^{(i)}, \ \mathbf{t}^{(i)}), \ \ \mathbf{z}^{(i)} = (\mathbf{y}^{(i)}, \mathbf{l}^{(i)}),$$

We will first learn a text summarization model, i.e., we seek to learn a discriminant function,

$$\mathcal{P}(\mathbf{t}, \mathbf{l}) : \mathcal{T} \times \mathcal{L} \longrightarrow \mathcal{R},$$

which measures the degree to which the subset $\mathbf{l}$ is a suitable summary for the textual information set $\mathbf{t}$.

Again, a feature vector $\Phi(\mathbf{t}, \mathbf{l})$ is employed to describe each pair $(\mathbf{t}, \mathbf{l})$. The discriminant function $\mathcal{P}(\mathbf{t}, \mathbf{l})$ is assumed to be linear in the feature vector $\Phi(\mathbf{t}, \mathbf{l})$, which can be expressed as follows:

$$\mathcal{P}(\mathbf{t}, \mathbf{l}) = \bar{\mathbf{w}}^T \Phi(\mathbf{t}, \mathbf{l}). \qquad (6)$$

To transfer the textual data into knowledge on summarization based on video features only, we emphasize the similar distribution of video features and text features by introducing the following additional constraints.

New Constraint:

$$\begin{aligned} &\mathbf{w}^T \Psi(\mathbf{v}^{(i)}, \mathbf{y}^{(i)}) \geq \mathbf{w}^T \Psi(\mathbf{v}^{(i)}, \mathbf{y}) \\ &+ (\bar{\mathbf{w}}^T \Phi(\mathbf{t}^{(i)}, \mathbf{l}^{(i)}) - \bar{\mathbf{w}}^T \Phi(\mathbf{t}^{(i)}, \mathbf{l})) - \xi_i, \end{aligned} \qquad (7)$$

where $\bar{\mathbf{w}}$ is obtained from a good text summarization model mentioned above.

According to our assumption that $\Psi(\mathbf{v}, \mathbf{y})$ and $\Phi(\mathbf{t}, \mathbf{l})$ are closely related, it is believed that $\bar{\mathbf{w}}^T \Phi(\mathbf{t}, \mathbf{l})$ can better represent the difference between candidate montages $\mathbf{y}$, as long as we employ a relatively accurate text summarization model, in other word, a suitably estimated $\bar{\mathbf{w}}^T$. In fact, the text features make $\bar{\mathbf{w}}^T \Phi(\mathbf{t}^{(i)}, \mathbf{l}^{(i)}) - \bar{\mathbf{w}}^T \Phi(\mathbf{t}^{(i)}, \mathbf{l})$ a word-level metric, which is generally better than shot-level metric. In this way, the text information is transferred to learn a more suitable weight vector for the video feature vector $\Psi(\mathbf{v}, \mathbf{y})$.

In fact, the above constraints can also be applied to a much wider situation. Suppose there exists a one-to-one mapping between units in source and target domains. Then $\bar{\mathbf{w}}^T \Phi(\mathbf{t}, \mathbf{l})$ can provide another measure for the difference between candidates $\mathbf{y}$. Thus these constraints will benefit the model training in the target domain, especially when the task is easier in the source domain.

4.2 Discussion of Solutions for Different Types of Video

For different types of video, their relation with the associated textual information can vary a lot. TV series, of course, are known to be also closely related to the transcripts. However, for other types of videos, the corresponding textual information does not necessarily have a close relation with the videos. Take home movies as an example, a considerable part of the associated text are usually comments about

the whole event, thus lacking the important connection with the details described in a single shot. For field sports, textual information frequently describes the background, history, gossip, and so on, thus also fail to reflect the content of corresponding shot. Under this situation, it may be a good idea to also incorporate some measurement of the strength of the relation, so as to give a higher weight to the transfer of the more correct knowledge.

Using $r(\mathbf{y}, \mathbf{l})$ to denote the degree of fitness of one subset of shots with the corresponding textual information, the following constraints can be used instead:

$$\mathbf{w}^T\Psi(\mathbf{v}^{(i)}, \mathbf{y}^{(i)}) \geq \mathbf{w}^T\Psi(\mathbf{v}^{(i)}, \mathbf{y}) + (\bar{\mathbf{w}}^T r(\mathbf{y}^{(i)}, \mathbf{l}^{(i)})\Phi(\mathbf{t}^{(i)}, \mathbf{l}^{(i)}) - \bar{\mathbf{w}}^T r(\mathbf{y}, \mathbf{l})\Phi(\mathbf{t}^{(i)}, \mathbf{l})) - \xi_i,$$

In our experiment, we just focus on dealing with TV series, the implementation of the above type of constraints is left to future work.

4.2.1 Combined Optimization Problem

Employing the additional constraints we discussed, we propose to train a summarization model using both text and video features through the following optimization problem:

Optimization Problem 2.

$$\min_{\mathbf{w}, \xi \geq 0} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i, \tag{8}$$

subjected to:

$$\forall i, \forall \mathbf{y} \in \mathcal{Y} \setminus \mathbf{y}^{(i)} : \xi_i \geq 0,$$

$$1)\ \mathbf{w}^T\Psi(\mathbf{v}^{(i)}, \mathbf{y}^{(i)}) \geq \mathbf{w}^T\Psi(\mathbf{v}^{(i)}, \mathbf{y}) + \Delta(\mathbf{y}^{(i)}, \mathbf{y}) - \xi_i,$$

$$2)\ \mathbf{w}^T\Psi(\mathbf{v}^{(i)}, \mathbf{y}^{(i)}) \geq \mathbf{w}^T\Psi(\mathbf{v}^{(i)}, \mathbf{y}) + (\bar{\mathbf{w}}^T\Phi(\mathbf{t}^{(i)}, \mathbf{l}^{(i)}) - \bar{\mathbf{w}}^T\Phi(\mathbf{t}^{(i)}, \mathbf{l})) - \xi_i.$$

5. SUMMARIZATION THROUGH STRUCTURED LEARNING

The space $\mathcal{Y}$ of all possible subsets is complex. We solve the optimization problem (8) following the general cutting plane algorithm [30, 35].

5.1 Learning Algorithm

The space $\mathcal{Y}$ of all possible subsets is complex. In order to solve the optimization problem defined in Equation (8) efficiently, we employed the cutting plane algorithm [30, 35]. It iteratively adds constraints until the problem has been solved with a desired tolerance ϵ. We start with a group of empty working sets $\mathbf{y}_i$, $\mathbf{y}_i^{'}$, for $i = 1, \ldots, n$. Then, we iteratively find the most violated constraints $\bar{\mathbf{y}}$, $\bar{\mathbf{y}}^{'}$, for each $(\mathbf{v}^{(i)}, \mathbf{y}^{(i)})$ corresponding to the two constraints in Equation (8), respectively. They are then added to the corresponding working sets, and $\mathbf{w}$ is updated with respect to the new combined working set. The learning algorithm is presented in Algorithm 1. It is guaranteed to halt within a polynomial number of iterations [30].

For each iteration, we need to solve

$$\operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}} \mathcal{P}(\mathbf{v}^{(i)}, \mathbf{y}^{(i)}, \mathbf{y}) \equiv \operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}} \Omega(\mathbf{y}^{(i)}, \mathbf{y}) + \mathbf{w}^T\Psi(\mathbf{v}^{(i)}, \mathbf{y})$$

for $\Omega(\mathbf{y}^{(i)}, \mathbf{y}) = \Delta(\mathbf{y}^{(i)}, \mathbf{y})$, or $\bar{\mathbf{w}}^T\Phi(\mathbf{t}^{(i)}, \mathbf{l}^{(i)}) - \bar{\mathbf{w}}^T\Phi(\mathbf{t}^{(i)}, \mathbf{l})$. A greedy algorithm, described in Algorithm 2, is proposed to solve this problem where we repeatedly select the shot y satisfying the condition that $\bar{\mathbf{y}} \cup \{y\}$ is the shot set with the highest score. The algorithm ends with an extracted shot set of size k. This algorithm has the same approximation bound as the greedy algorithm proposed by Khuller et al. [13] to solve the budgeted maximum coverage problem, that is to say, a $(1 - \frac{1}{e})$-approximation bound. According to [13, 30, 35], Algorithm 1 has a polynomial time complexity overall.

Algorithm 1 Cutting plane algorithm

```
Input (v^(1), y^(1)), ..., (v^(n), y^(n)), c > 0, ε > 0
y_i = ∅, y_i' = ∅, y_i'' = ∅  for i = 1, ..., n
repeat
  for i = 1, 2, ..., n do
    ω ≡ w^T(Ψ(v^(i), y^(i)) − Ψ(v^(i), y))
    H(y) = Δ(y^(i), y) − ω
    H'(y) = (w̄^T Φ(t^(i), l^(i)) − w̄^T Φ(t^(i), l)) − ω
    Compute: ȳ = argmax_y H(y), ȳ' = argmax_y H'(y)
    Compute actual slack: ξ_i = max{0, max_{y∈W_i} H(y),
      max_{y∈y_i'} H'(y)}
    if (H(ȳ) > ξ_i + ε) or (H'(ȳ) > ξ_i + ε) then
      Add constraint to working set y_i ← y_i ∪ {ȳ},
      y_i' ← y_i' ∪ {ȳ'}
      w ← Optimize over ∪_i(y_i ∪ y_i')
    end if
  end for
until no working set has changed during iteration.
```

Algorithm 2 Greedy algorithm for shot subset selection

```
Input: v, y
Initialize prediction ȳ ← ∅
for i = 1, 2, ..., k do
  y ← argmax_{y∉ȳ} P(v, y, ȳ ∪ {y})
end for
return ȳ
```

5.2 Making Prediction

According to (1) and (2), we generate the summary for a given new video $\mathbf{v}$ using:

$$\operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}} \mathcal{P}(\mathbf{v}, \emptyset, \mathbf{y}) \equiv \operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}} \mathcal{F}(\mathbf{v}, \mathbf{y}) = \operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}} \mathbf{w}^T\Psi(\mathbf{v}, \mathbf{y}),$$

which is a special case of Algorithm 2.

6. EXPERIMENTAL RESULTS

6.1 Data Set

Our training and testing data are prepared as follows: For each episode, we collect the following data: the episode itself, the corresponding transcript, the preview, and the drama. In our experiments, the preview is employed as the summary of its corresponding episode, and the drama is used as the ground-truth summary of the transcript. The data set in our experiments contains 100 episodes derived from several popular U.S. TV shows [3] with the statistics presented in Table 1. We perform 5-fold cross validation to determine the parameters in the experiments. The reported performance is averaged over all 5-folds.

[3]One example is from fox.com/24.

Table 1: Summary of our data set

	Dataset
Number of Episodes	100
Average Episode Length	40 min 39 s
Average Preview Length	1 min 58 s
Resolution	640 × 352

In our experiment, the video features $\Psi(\mathbf{v}, \mathbf{y})$ are generally keyframe-based. IBM's IMARS [1] is used to split single video into shots and extract one keyframe for each shot. For each keyframe, five kinds of features, including lab histogram [18], Law [18], human color vision [18], GIST [22] and SIFT [17], are extracted, leading to a feature vector with 1514 dimensions. We use the following five kinds of text features $\Phi(\mathbf{t}, \mathbf{l})$: word frequency, position, thematic word, length and n-gram [15], leading to a feature vector with 2090 dimensions.

6.2 Performance Evaluation

We apply two measures, F_1 and *Simi* to evaluate the performance. The metrics provides by TRECVID is not completely employed, as they are not focus on summarization measurement.

6.2.1 F_1 Evaluation

F_1 measurement is widely used in summarization evaluation. In F_1 metric, the predicted summary $\bar{\mathbf{y}}$ and the ground-truth summary $\mathbf{y}$ are compared directly and the precision, recall, F_1 scores are calculated as follows:

$$F_1(\mathbf{y}, \bar{\mathbf{y}}) = \frac{2pr}{p+r}, \quad p = \frac{< \mathbf{y}, \bar{\mathbf{y}} >}{< \bar{\mathbf{y}}, \bar{\mathbf{y}} >}, \quad r = \frac{< \mathbf{y}, \bar{\mathbf{y}} >}{< \mathbf{y}, \mathbf{y} >}.$$

6.2.2 *Simi* Evaluation

To give a more detailed and precise evaluation of the performance, we measure the difference between the ground-truth video summary and candidate video summary through the similarities between their shots. Let $vsmi(v_i, v_j)$ to denote the similarity of two shots v_i and v_j, which is determined by the similarities of their corresponding keyframes. Suppose the corresponding keyframes of v_i and v_j are denoted as $\{k_{i1}, k_{i2}, \ldots, k_{im}\}$ and $\{k_{j1}, k_{j2}, \ldots, k_{jn}\}$, respectively, we calculate $vsmi(v_i, v_j)$ through:

$$vsmi(v_i, v_j) = \frac{1}{m} \sum_{i=1}^{m} \max_{j=1,\ldots,n} psmi(k_i, k_j),$$

where $psmi(k_i, k_j)$ denotes the similarity of two keyframes, measured by a widely used image comparison tool ImageMagick [4] based on the pixel difference.

In order to compare the predicted summary $\bar{\mathbf{y}} = \{\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_{k'}\}$ and the ground-truth summary $\mathbf{y} = \{v_1, v_2, \ldots, v_k\}$, we define $Simi(\mathbf{y}, \bar{\mathbf{y}})$ as follows:

$$Simi(\mathbf{y}, \bar{\mathbf{y}}) = \frac{1}{k} \sum_{i=1}^{k} \max_{j=1,\ldots,k'} vsmi(v_i, \bar{v}_j)$$

6.3 Overall Performance

In this section, we list several supervised and unsupervised methods that are used for comparison in our experiments.

[4] www.imagemagick.org

6.3.1 Algorithms for Comparison

We choose two approaches based on video data only and three approaches based on both video and textual data as our baselines.

`SVM: SVM` is widely used to train a binary classifier. We use `SVM` to classify summary shots and non-summary shots based on video features only.

`Clustering [31]:` It builds a graph based on the similarities between extracted keyframes. A time-constrained clustering algorithm is proposed to classify all shots into groups, and then select the most representative one of each cluster to form montages.

`Mixed-Clustering [36]: Mixed-Clustering` makes a further step over the above one by calculating the mixed similarity scores based on both video and textual data.

`KP-Transcript [26]:` This approach views the video summarization task as a knapsack problem. It aims at maximizing segment score derived from corresponding transcripts and constrained by segment duration.

`Mixed-Str-SVM: Mixed-Str-SVM` shares a similar structured learning framework with our approach, but with no constraints for knowledge transfer involved. It makes use of both video and text features in montage prediction.

`Trans-Str-SVM:` This is our proposed approach.

6.3.2 Results and Analysis

Table 2 gives the performance of our approach and the baselines.

As shown in Table 2, `Clustering` gives the worst performance. We think this is due to the following reasons: First, it only utilizes on the visual similarities between different shots. Second, current approaches can hardly give an accurate measure for the similarities between shots or keyframes based on video or image features only. As a result, it can not obtain accurate summaries since two shots with similar colors can actually tell quite different stories. `SVM` performs a little better than `Clustering`, which demonstrates the advantage of supervised learning. Moreover, it proves that it is quite difficult to distinguish summary shots from non-summary shots.

Approaches that utilize both video and text features perform notably better than with those approaches using only video features. For example, `Mixed-Clustering` results in a remarkable improvement over the performance of `Clustering`, thereby shows the great benefits brought by textual data. In particular we may conclude that the shots similarity may be better measured by taking text information into consideration, since main themes are usually easier to extract from textual data. `KP-Transcript` confirms that it is reasonable to view subset selection problems as knapsack problems by obtaining much better result. Notice that in this method, text features actually play a major role in determining which shots are to be selected as summary for the given video, while the video features are mainly used for constraint generation. `Mixed-Str-SVM` performs the best among all these approaches. The improvement shows the effectiveness of structured learning, especially in dealing with textual data, and provides the possibility for transferring text knowledge for video summarization under our framework.

Our method outperforms the two methods based on video features only, achieves an improvement of 33.4% and 19.0% over `Clustering` in terms of F_1 and *Simi* metrics, respec-

Table 2: Performance comparison between our approach and several baselines

	SVM	Clustering	Mixed-Clustering	KP-Transcript	Mixed-Str-SVM	Trans-Str-SVM
F_1	0.2226	0.2059	0.3799	0.3999	0.4264	0.2746
Simi	0.2763	0.2537	0.3983	0.4332	0.4550	0.3020

Table 3: Performance of different models measured by F_1

	Str-SVM	Trans-Str-SVM	Str-SVM(G)	Trans-Str-SVM(G)	Str-SVM(B)	Trans-Str-SVM(B)
Fold1	0.1916	0.2362	0.2047	0.2299	0.1609	0.2510
Fold2	0.1718	0.1772	0.1741	0.1770	0.1666	0.1776
Fold3	0.2903	0.3267	0.2813	0.3530	0.3409	0.1773
Fold4	0.2415	0.3006	0.2402	0.3024	0.2667	0.2667
Fold5	0.2815	0.3214	0.2603	0.3239	0.3667	0.3111
Total	0.2354	0.2724	0.2352	0.2828	0.2360	0.2307

The additional tag (G) denotes the results on test cases easy for text summarization, (B) denotes the results on others. Table 4 uses same tags.

Table 4: Performance of different models measured by *Simi*

	Str-SVM	Trans-Str-SVM	Str-SVM(G)	Trans-Str-SVM(G)	Str-SVM(B)	Trans-Str-SVM(B)
Fold1	0.2328	0.2724	0.2470	0.2707	0.1999	0.2763
Fold2	0.2203	0.2097	0.2207	0.2065	0.2193	0.2174
Fold3	0.3290	0.3524	0.3206	0.3776	0.3770	0.2097
Fold4	0.2768	0.3302	0.2767	0.3330	0.2800	0.2767
Fold5	0.3011	0.3477	0.2798	0.3520	0.3862	0.3303
Total	0.2720	0.3025	0.2716	0.3132	0.2735	0.2594

tively. Moreover, it outperforms SVM by 23.4% and 9.3% in term of F_1 and *Simi*, respectively. This also demonstrates that text information is useful for video summarization. On the other hand, compared with our approach, there is a considerable performance gain for Mixed-Clustering, KP-Transcript, and Mixed-Str-SVM. This can be attributed to the fact that text information are available in the test phrase for these methods, while in our setting no text data is available in the test phrase. Therefore, they can be viewed as the upper-bound of the proposed method. We include these methods for a comprehensive evaluation.

According to the above results, we can also find that F_1 and *Simi* show the similar trend in performance measurement. Therefore, in the following experiments, the performance is only measured by F_1 metric.

6.4 Performance of Transfer Learning

6.4.1 General Performance

To identify how the knowledge in the text domain help our video summarization task, we provide two models through a strategy selection based on our approach. We use Str-SVM and Trans-Str-SVM to denote the model trained with only video features and the model trained with both video and text features, respectively. To better understand how text information can aid video summarization, we divide the data set into two parts based on the performance of the text summarization model on each test case. Specifically, the boundary is determined by an empirical threshold under F_1 metric. The results on those two data sets are also presented, and (G) denotes the result on test cases for which text summary can be easily obtained while (B) denotes the result for the rest.

According to Table 3 and Table 4, we find that Trans-Str-SVM gains the advantage over Str-SVM in all five folds measured by F_1, and also outperforms Str-SVM in most cases measured by *Simi*. Trans-Str-SVM improve the performance over Str-SVM by 15.7% and 11.2% in average measured by F_1 and *Simi*, respectively. Further more, there is a relatively smaller increase in terms of *Simi*, this may due to the fact that the *Simi* metric emphasizes the similarities between keyframes. However, image similarity can not accurately describe the effective structures in the summarization task. The transfer of text knowledge, on the other hand, pay more attention to the structure of video.

The quality of textual summary also has a influence on the performance of Trans-Str-SVM. For those episodes where text summarization model performs relatively well, Trans-Str-SVM generally achieves a much greater performance gain over Str-SVM than the average case, makes an improvement of 20.2% and 15.3% in terms of F_1 and *Simi* metrics, respectively. For those episodes difficult for text summarization, Trans-Str-SVM obtains comparable or a litte worse performance than Str-SVM. As a result, we believe that useful knowledge can be transferred through the proposed constraints when text summaries can be obtained accurately. To a large extent, its accuracy determines how much benefits we can get from knowledge transfer. This also demonstrates that video features and corresponding text features are highly correlated.

According to the above results, we can now come to the conclusion that the constraints generated from textual data can substantially contribute to video summarization. They indeed transfer text information into knowledge on how to summarize based on video features only.

6.4.2 Case Study

To further demonstrate the significant benefits brought by knowledge transferred from textual data, we select one test case and compare the performance of Str-SVM and

Table 5: Performance of our approach employing different text summarization models

	Good-Trans1	Good-Trans2	Bad-Trans1	Bad-Trans2
F_1	0.2711	0.2746	0.2653	0.2657
Simi	0.3018	0.3020	0.2936	0.2946

Trans-Str-SVM manually. The keyframes of ground-truth montage, keyframes of montage predicted by Str-SVM, and keyframes of montage predicted by Trans-Str-SVM are presented in Figure 4.

From Figure 4 we can find that Str-SVM tends to extract the keyframes owning similar hue with the ground-truth summary. They may involve no human figures that share similar backgrounds with the keyframes in the ground-truth summary, or present several scenes holding diversity colors or stripes. However, those keyframes usually are not the right answer, as video summary prefers keyframes with single scene and person.

Trans-Str-SVM may fail to learn the color diversity presented by the Str-SVM. However, it extracts more accurate keyframes. This can be attributed to the structure information it learns from knowledge transfer. Keyframes extracted by Trans-Str-SVM are generally in consistent with the previous mentioned principles of video summary(video summary prefers keyframes with single scene and person). We believe that those principles are closely related with the video's structures.

6.5 Performance of Text Summarization Model

In this section, we test the influence of text summarization model on the performance of our approach. Four video summarization models are trained separately on different training sets or using different parameters in model training. We denote Good-Trans1, Good-Trans2 to be the text summarization model with better performance in text summarization task, and Bad-Trans1, Bad-Trans2 to be the model getting a relatively worse result. Table 5 presents the performance of our approaches employing the above four models.

According to Table 5, the model Good-Trans1 and Good-Trans2 make an average increase of 2.7% and 2.7% over the model Bad-Trans1 and Bad-Trans2 measured by F_1 and *Simi*, respectively. This illustrates that transferring text information in a correct manner is very important. When using unreliable text summarization models, the corresponding constraints may provide an inaccurate measurement of the difference between candidate montages. In fact, this can play quite a negative effect on adjusting weighting for those video features. Thus we can get a worse video summarization model.

7. CONCLUSIONS AND FUTURE WORK

In this paper, a novel approach is proposed to make use of textual data to help the video summarization task using only video data. In order to transfer text information to help the summarization based on video features only, we construct a transferable structured learning framework to train a video summarization model with a set of constraints generated from textual data, which is intended for knowledge transfer. In particular, a text summarization model is firstly trained and then used to generate constraints which transfer the knowledge from textual summarization and give a more detailed and precise measurement of the differences among shots. Experimental results demonstrate the effectiveness of our approach. The performance of our method achieves a remarkable improvement over a set of state-of-the-art supervised and unsupervised methods which indicates that the constraints can transfer knowledge from the textual data effectively.

In future work, we plan to transfer text information to knowledge for video summarization in other ways. For example, a text-video dictionary can be employed or generated from web resources such as Flickr, to transfer text features into video features. Moreover, state-of-the-art video tagging approaches can be employed to generate textual data from videos or keyframes. The text features generated from those data can be ensured to be more consistent with corresponding video features, which can be expect to benefit video summarization a lot.

8. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable and constructive comments. Gui-Rong Xue thanks the support of NSFC project (No. 60873211), RGC/NSFC project (No. 60910123) and Open Project Program of the state key lab of CAD & CG (No. A0801), Zhejiang University. Part of the work was supported by NSF Grant IIS-1049694.

9. REFERENCES

[1] Ibm multimedia analysis and retrieval system, http://www.alphaworks.ibm.com/tech/imars.

[2] Trec video, http://trecvid.nist.gov/.

[3] A. Bagga, J. Hu, J. Zhong, and G. Ramesh. Multi-source combined-media video tracking for summarization. *Pattern Recognition, International Conference on*, 2:20818, 2002.

[4] B.-W. Chen, J.-C. Wang, and J.-F. Wang. A novel video summarization based on mining the story-structure and semantic relations among concept entities. *Multimedia*, 11(2):295–312, Feb. 2009.

[5] G. Cohen, A. Amir, D. Ponceleon, B. Blanchard, D. Petkovic, and S. Srinivasan. Using audio time scale modification for video browsing. In *HICSS*, page 3046, Washington, DC, USA, 2000. IEEE Computer Society.

[6] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu. Translated learning: Transfer learning across different feature spaces. In *NIPS*, pages 353–360, 2008.

[7] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu. Transferring naive bayes classifiers for text classification. In *AAAI*, pages 540–545. AAAI Press, 2007.

[8] M. Detyniecki and C. Marsala. Video rushes summarization by adaptive acceleration and stacking of shots. In *TVS*, pages 65–69, New York, NY, USA, 2007. ACM.

[9] L. He, E. Sanocki, A. Gupta, and J. Grudin. Auto-summarization of audio-video presentations. In *MULTIMEDIA*, pages 489–498, New York, NY, USA, 1999. ACM.

[10] Q. Huang, Z. Liu, A. Rosenberg, D. Gibbo, and B. Shahraray. Automated generation of news content hierarchy by integrating audio, video, and text information. In *International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[11] J. Jiang and C. Zhai. Instance weighting for domain

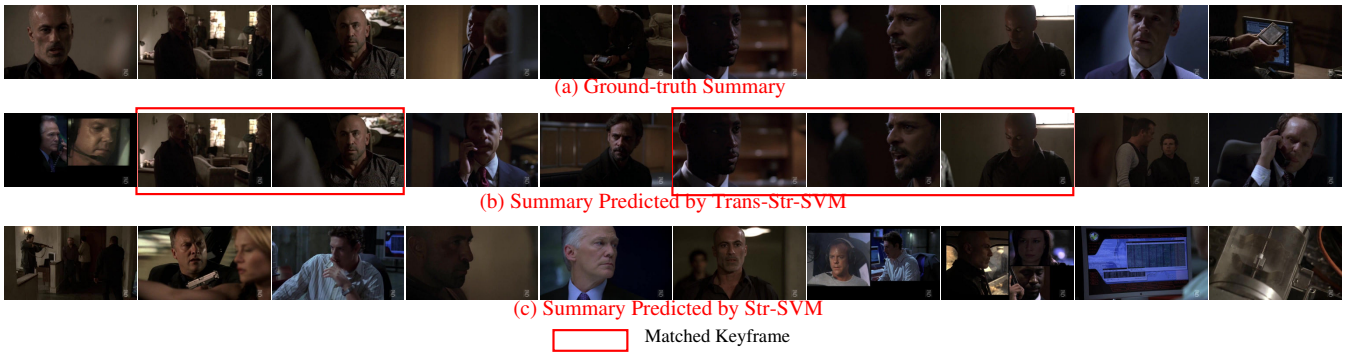

Figure 4: A example of our video summarization task

adaptation in nlp. In *ACL*. The Association for Computer Linguistics, 2007.

[12] S. X. Ju, M. J. Black, S. Minneman, and D. Kimber. Summarization of video-taped presentations: Automatic analysis of motion and gesture. *IEEE Trans. on Circuits and Systems for Video Technology*, 8:686–696, 1998.

[13] S. Khuller, A. Moss, and J. Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45, 1999.

[14] N. D. Lawrence and J. C. Platt. Learning to learn with the informative vector machine. In *ICML*, page 65, New York, NY, USA, 2004. ACM.

[15] L. Li, K. Zhou, G.-R. Xue, H. Zha, and Y. Yu. Enhancing diversity, coverage and balance for summarization through structure learning. In *WWW*, pages 71–80, New York, NY, USA, 2009. ACM.

[16] Y. Li, S.-H. Lee, C.-H. Yeh, and C. C. J. Kuo. Techniques for movie content analysis and skimming: tutorial and overview on video abstraction techniques. *Signal Processing Magazine*, 23(2):79–89, 2006.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[18] R. B. M. Sonka, V. Hlavac. *Image Processing, Analysis, and machine vision*. 2007.

[19] L. Mihalkova, T. Huynh, and R. J. Mooney. Mapping and revising markov logic networks for transfer learning. In *AAAI*, pages 608–614. AAAI Press, 2007.

[20] M. Mills, J. Cohen, and Y. Y. Wong. A magnifier tool for video data. In *CHI*, pages 93–98, New York, NY, USA, 1992. ACM.

[21] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang. Video summarization and scene detection by graph modeling. *IEEE transactions on circuits and systems for video technology*, 15(2):296–305, 2005.

[22] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, 2001.

[23] D. M. Russell. A design pattern-based video summarization technique: Moving from low-level signals to high-level structure. *Hawaii International Conference on System Sciences*, 3:3048, 2000.

[24] M. A. Smith and T. Kanade. Video skimming and characterization through the combination of image and language understanding techniques. pages 370–382, 2001.

[25] Y. Taniguchi, A. Akutsu, Y. Tonomura, and H. Hamada. An intuitive and efficient access interface to real-time incoming video based on automatic indexing. In *MULTIMEDIA*, pages 25–33, New York, NY, USA, 1995. ACM.

[26] C. M. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon, and E. J. Delp. Automated video program summarization using speech transcripts. *Multimedia*, 8(4):775–791, 2006.

[27] M. E. Taylor and P. Stone. Cross-domain transfer for reinforcement learning. In *ICML*, pages 879–886, New York, NY, USA, 2007. ACM.

[28] R. Tibshirani and G. Hinton. Coaching variables for regression and classification. *Statistics and Computing*, 8(1):25–33, 1998.

[29] B. T. Truong and S. Venkatesh. Generating comprehensible summaries of rushes sequences based on robust feature matching. In *TVS*, pages 30–34, New York, NY, USA, 2007. ACM.

[30] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 6:1453–1484, 2005.

[31] S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky. Video manga: generating semantically meaningful video summaries. In *MULTIMEDIA*, pages 383–392, New York, NY, USA, 1999. ACM.

[32] Z. Xiong, R. Radhakrishan, A. Divakaran, and Y. Ishikawa. Generation of sports highlights using motion activity in combination with a common audio feature extraction framework. In *ICIP*, 2003.

[33] M. M. Yeung and B. L. Yeo. Video visualization for compact presentation and fast browsing of pictorial content. *IEEE Trans. on Circuits and Systems for Video Technology*, 7:771–785, 1997.

[34] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *ICML*, pages 1169–1176, New York, NY, USA, 2009. ACM.

[35] Y. Yue and T. Joachims. Predicting diverse subsets using structural svms. In *ICML*, pages 1224–1231, New York, NY, USA, 2008. ACM.

[36] X. Zhu, J. Fan, A. K. Elmagarmid, and X. Wu. Hierarchical video content description and summarization using unified semantic and visual similarity. *Multimedia Syst.*, 9(1):31–53, 2003.