JIS

Automatic categorization and summarization of documentaries

Kezban Demirtas and Nihan Kesim Cicekli

Department of Computer Engineering, METU, Ankara, Turkey

Ilyas Cicekli

Department of Computer Engineering, Bilkent University, Ankara, Turkey

Abstract.

In this paper, we propose automatic categorization and summarization of documentaries using subtitles of videos. We propose two methods for video categorization. The first makes unsupervised categorization by applying natural language processing techniques on video subtitles and uses the WordNet lexical database and WordNet domains. The second has the same extraction steps but uses a learning module to categorize. Experiments with documentary videos give promising results in discovering the correct categories of videos. We also propose a video summarization method using the subtitles of videos and text summarization techniques. Significant sentences in the subtitles of a video are identified using these techniques and a video summary is then composed by finding the video parts corresponding to these summary sentences.

Keywords: video categorization; video summarization; text summarization; WordNet domains

1. Introduction

Video content is being used in a wide number of domains ranging from commerce to security, education and entertainment. People want to search and find this content according to its semantics. Creating searchable video archives had become an important requirement for different domains as a result of the increase in the amount of multimedia content. Narrowing down the user's search space by categorizing videos can help people to solve this problem. Since there is a huge amount of videos to categorize, automatic categorization is an important research area [1–4]. Video summarization helps people to decide whether they really want to watch a video. Summarization algorithms present condensed versions of a full length video by identifying the most significant parts. In order to have an idea about the content of a video, using such a summary is much easier than going through all of the footage.

Correspondence to: Ilyas Cicekli, Deptartment of Computer Engineering, Bilkent University, 06800 Bilkent, Ankara, Turkey. Email: ilyas@cs.bilkent.edu.tr

In video categorization, video is generally classified into one of several broad categories such as documentary type (e.g. geography, animals religion) or movie genre (e.g. action, comedy, drama, horror). In order to classify videos automatically, features are drawn from three modalities: visual, audio or text. Also some combinations of these features can be exploited together. Therefore video classification approaches could be divided into four groups: text-based approaches, audio-based approaches, visual-based approaches and those that use some combination of visual, audio and text features. Text-based approaches [2, 4–6] are the least common in the video classification literature but have several benefits over other approaches. First of all, text processing is a more lightweight process than video and audio processing. Also text categorization techniques have been studied extensively in the computational linguistics literature [1, 3]. This accumulation can be exploited in video classification domain. Beside this, the human language in a video carries more semantic information than its visual/audio features. Words have meaning to humans and some tend to be associated with certain categories. Another benefit of using text features is that, by using some lexicon, such as WordNet, concept learning can be performed.

Video summaries are either used individually or integrated into various applications, such as browsing and searching systems. There are two main trends in video summarization: still image summaries and moving image summaries. The former are based on extracting individual key frames representing the content of the video in a static way [7]. Generally video is segmented into shots and the key frames representing these shots are selected to be included in the summary. The latter are a collection of original video parts [8, 9]. These summaries can be classified into two subtypes: previews and summaries. Video previews present the most interesting parts of a video, for example a movie trailer, whereas video summaries keep the semantic meaning of the original. Since video has a multimodal nature, summarization can be performed by using the image features, audio features or text features of video. A combination of these can be exploited together.

In this paper we propose to use automatic categorization and summarization techniques in one framework to help users first find the category of the video and then present its summary so that they can decide easily whether the selected video is of any interest to them. We aim to use text information only in order to determine how the data associated with the video are helpful in searching the semantic content of videos. For this purpose, we have chosen documentary videos as the application domain as the speech usually consists of a monologue and it mentions the things seen on the screen. The subtitles provide the speech content with the time information which is used to retrieve the relevant video pieces.

Two methods for video categorization, both based on text processing, are proposed. The first, category label assignment, makes categorization by applying natural language processing techniques on video subtitles and uses the WordNet lexical database and WordNet domains. The method is based on an existing video categorization algorithm [6] and makes some extensions to this. The TextRank algorithm [10] is used for keyword selection and one third of words are selected as keywords. In our implementation, we do not use this keyword rate but instead determine the number of keywords experimentally. Additionally, our algorithm makes use of the title of a documentary video in addition to the subtitles. The title gives important clues about the type of video because generally they are selected in order to reflect the content of the documentary. For example, the category of the documentary 'War of the century' is 'War', or the category of the documentary 'Planet Earth – mountains' is 'Geography'.

The second method, categorization by learning, has the same steps for extracting WordNet domains but performs categorization by using a learning module which learns the general WordNet domain distributions of categories. When categorizing a video, its WordNet domain distribution is analysed and the most similar category is assigned.

For automatic video summarization, we make use of two text summarization algorithms [10, 11] and combine the results to constitute a summary. Text summarization techniques identify the significant parts of a text to constitute a summary. We extract a summary of video subtitles and then the corresponding video parts are found. By combining the video parts, we create a moving image summary of the original. In our summarization approach, we take the advantage of the documentary

video characteristics. For example, in a documentary about 'animals', when an animal is seen on the screen, the speaker usually mentions that animal. So, when we find the video parts corresponding to the summary sentences of a video, those video parts are closely related with the summary sentences. Hence we obtain a semantic video summary giving the important parts of a video.

The rest of the paper is organized as follows. In Section 2, we discuss the related work in video categorization and video summarization. Section 3 describes our algorithms for video categorization and presents an evaluation of the algorithms. In Section 4, we give the description of our video summarization approaches and an evaluation of these approaches. Finally in Section 5, conclusions and possible future work are discussed.

2. Related work

In this section, we discuss the related work in video categorization and video summarization. We performed the latter by utilizing text summarization techniques and therefore a summary of the literature on both is presented.

2.1. Related work in video categorization

Video categorization algorithms assign a meaningful label to a video such as 'sports video' or 'comedy video'. The required features are drawn from three modalities: visual, audio and text. So, video categorization approaches can be classified as visual-based, audio-based and text-based. Some approaches use a combination of these three features.

Since the main topic of this paper is the categorization of videos using text features, we present here the related work on video categorization based on text processing. The text associated with a video can be viewable text or a transcript of the dialogue. The former is the text placed on the screen and some optical character recognition (OCR) methods should be used in order to use this. The latter can be provided in the form of closed/open captions or subtitles. Alternatively, it can be obtained by using speech recognition methods.

Zhu et al. [4] performed automatic news video story categorization based on the closed-captioned text. They segmented news video into stories using the demarcations which indicate the topic changes in the text. Then for each story, a category is assigned by extracting a list of keywords and further processing them.

Brezeale and Cook [1] used text (closed captions) and visual features separately in video classification. To classify a movie, the closed captions are firstly extracted and stop words are subsequently removed. Each word is then stemmed by removing the suffixes to find the root. By using these stemmed words, a term feature vector is generated. Classification is performed using a support vector machine (SVM). There are 15 genres of movies from the entertainment domain and the evaluation is performed on 81 movies.

Wang et al. [12] used text features for classification purposes. News videos were assigned to one of 10 categories and the spoken text was extracted using speech recognition methods. Text derived from speech recognition, however, generally has a fairly high error rate.

Qi et al. [13] classified a news video into types of news stories. First, the shots and, if necessary, scenes of video were detected using audio and visual features. The closed captions and scene text detected by the OCR methods were then used for classification.

Katsiouli et al. [6] used subtitles for documentary classification. They performed categorization by using the WordNet lexical database and WordNet Domains [14] and applied natural language processing techniques on subtitles. They predefined documentary categories as geography, history, animals, politics, religion, sports, music, accidents, art, science, transportation, technology, people and war. Their categorization approach has achieved 69.4% accuracy. In this paper, a similar approach is followed with a different categorization algorithm and better results are obtained.

2.2. Related work in video summarization

In the literature, there are several approaches using the image, audio or text features in video summarization. Also some approaches use a combination of these features [1, 15]. Image features include changes in colour, texture, shape and motion of objects generated by the image stream of the video. By using these features, the shots of a video can be identified, such as cuts or fades. Cuts are represented by sharp changes while fades are identified by slower changes in image features. For instance, Ekin et al. [16] observed that the important scenes of a football game conform to long, medium and close-up view shots and these are then used in their summarization system.

In addition to shots, specific objects and events can be identified and this information could improve summarization performance. Knowledge of content domain could be helpful in the identification of objects within the video (e.g. anchor person) and events (e.g. the news headlines). The techniques presented in [14, 17] analyse image features from the video stream, and are domain specific. The systems in [16, 18–20] use image features to identify representative key frames for inclusion in the video summary and all are non-domain specific.

Audio features associated with a video include speech, music, sounds and silence. These are used to select candidate segments to be included in a video summary and domain-specific knowledge can be used to enhance the summary success. For example, excited commentator speech and excited audience sounds may show a number of potential events such as the start of a free kick, penalty kick, foul or goal [5]. Rui et al. [21] analysed the speech track to find exciting segments and events, such as baseball hits in baseball videos.

Text features play an important role in video summarization as they contain detailed information about the content. Pickering et al. [22] used accompanying subtitles to summarize television news. They extracted news stories and provided a summary for each one by using lexical chain analysis. Tsoneva et al. [23] created automatic summaries for narrative videos using textual cues available in subtitles and scripts. They extracted features like keywords, main characters' names and presence, and according to these features they identified the most relevant moments of video for preserving the storyline. In our video summarization system, we extracted moving image summaries of documentaries using video subtitles and text summarization methods.

2.3. Related work in text summarization

Text summarization techniques can be useful in video summarization since some videos have text related to the content and the summary is therefore an important resource. Text summarization techniques investigate different clues that could be used to identify important topics and ideas of the text. The summarization methods can be classified by the clues that they use in summarization. Summaries can be created by selecting the first sentences of text and this simple technique gives very good results in news articles and scientific reports [24].

In text, to emphasize the importance of a sentence some phrases are used such as 'significantly' and 'in conclusion' – these phrases are called 'bonus phrases'. On the other hand, some phrases reflect the unimportance of a sentence such as 'hardly' and 'impossible' – these phrases are called 'stigma phrases'. In addition to cue phrases, some formatting features like bold words and headers could enhance the summarization performance. The systems in [2, 25] make use of both in their summarization systems.

Weighted vectors of TF*IDF (Term Frequency * Inverse Document Frequency) values can be used to represent sentences. The TF*IDF value takes advantage of word repetition in the text which is a lexical cohesion type. Radev et al. [26] used such weighted vectors to find the important sentences in a summarization task. The summarization system in [27] uses an algorithm which is similar to Google's Pagerank [28] in order to select the summary sentences. Mihalcea and Tarau [10] proposed a summarization algorithm named TextRank which also relies on the Pagerank algorithm and uses the word repetition feature.

Lexical chains, which are sets of related words, can also be used for modelling lexical cohesion. Barzilay and Elhadad [29] used lexical chains to extract summaries and achieved good results. Many lexical cohesion-based algorithms [11, 30–32] are developed following the Barzilay/Elhadad

algorithm. Silber and McCoy [32] proposed a summarizer based on lexical chains and tried to improve the running time of the lexical chaining algorithm. Chali and Kolla [33] used lexical chains and offered a different sentence selection approach. In Ercan and Cicekli [11] the lexical cohesion structure of the text is exploited to determine the importance of sentences. Their summarization algorithm constructs the lexical chains of a text and identifies topics from them. The text is segmented with respect to these topics and the most important sentences are selected from these segments.

3. Automatic video categorization

Two algorithms for video categorization are proposed: category label assignment and categorization by learning. The first is based on the algorithm presented in [6]. The algorithm is extended by adding video name processing and changing the way the number of keywords in subtitle processing is determined. With these extensions, better results are obtained. According to this algorithm, a video is assigned a category label using the WordNet lexical database and WordNet domains [34] and applying natural language processing techniques on subtitles. In the second algorithm, categorization is done by learning. A learning module is implemented, which can be trained by using the videos of known categories. The algorithm starts with the preprocessing steps of the first algorithm and the categorization is performed by the learning module.

The common preprocessing steps of the two algorithms are given in Section 3.1. The first video categorization algorithm is given in Section 3.2, the second video categorization algorithm is given in Section 3.3, and the evaluation of these algorithms is given in Section 3.4.

3.1. Extracting WordNet domains

Initially, WordNet domains of a video are extracted and are used in both of the proposed categorization algorithms. The overview of extracting WordNet domains is given in Figure 1. The method for extracting WordNet domains starts with 'text preprocessing'. In this step, the sentences in the subtitle file are split, the words in every sentence are tagged with part of speech (POS) tags and the stop words are removed. The processed text is given to a 'keywords extraction' module which finds the keywords of the given text. Since these may carry more than one meaning, the 'word sense disambiguation' module finds the correct sense by using an adaptation of the Lesk algorithm [35]. Then the 'WordNet domains extraction' module finds the WordNet domains of the keywords corresponding to their correct senses. This module uses WordNet domains and considers the effect of the video title on categorization. Since titles give important clues about the category, this information is taken into consideration. Hence we obtain the WordNet domains of the video.

In the text preprocessing step, a subtitle is processed to find its sentences and the types of the words in these sentences are determined. A sample subtitle file is shown in Figure 2. After the sentences are extracted, a POS tagger is applied to the words, which determines the word class of each one in the sentence. The Stanford Log-linear Part-Of-Speech Tagger [36] was used for this purpose. The assigned part of speech tags consist of coded abbreviations conforming to the scheme of the Penn Treebank [37], the linguistic corpus developed by the University of Pennsylvania. For example, 'JJ' means 'Adjective', 'NNS' means 'Noun, plural' and NN means 'Noun, singular or mass'. After POS tagging, stop words (ones that do not contribute to the meaning of the sentence, such as 'above', 'the' and 'her') are removed from the sentences since these carry no semantics.

In order to select the most important words in the subtitle file for classifying the video, a keyword selection algorithm, namely the TextRank [10], is used. This algorithm builds a graph representing the text and applies a ranking algorithm to the vertices of the graph. The words are added to the graph as vertices for keywords extraction. Two vertices are connected if they have a co-occurrence relation. Two vertices co-occur if they are within a window of maximum N words, where N can be set to a value from 2 to 10. In our implementation N is set to 2. After building the graph, a graph-based ranking algorithm, derived from the PageRank algorithm [28], is used in order to decide the importance of a vertex. The basic idea of the algorithm is 'voting': when a vertex links to another

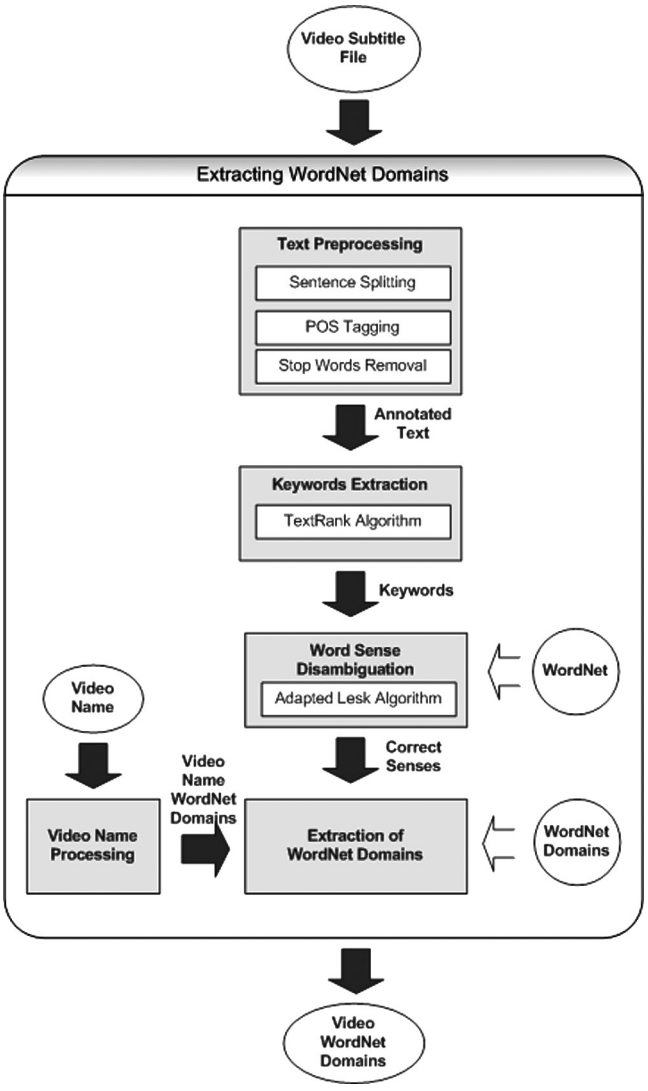

Fig. 1. Extracting WordNet domains.

one, it casts a vote for that vertex. Also, the importance of the vertex casting the vote determines the importance of the vote. Hence, the score of a vertex is computed by the votes that are cast for it and the score of the vertices casting these votes. Once the score of each vertex is computed, the vertices are sorted based on their scores and the top T vertices are selected as keywords. Generally, T is set to a third of the number of vertices in the graph. In our implementation, the number of the vertices selected as keywords is determined experimentally. Figure 3 gives a part of a subtitle file and the extracted keywords by the TextRank algorithm.

Word sense disambiguation (WSD) is the task of determining the correct sense of a word in a text. In order to find the correct senses of the keywords, we applied a WSD algorithm, which is presented in [35]. This algorithm is an adaptation of Lesk's dictionary-based algorithm. The adapted algorithm uses WordNet to include the glosses of the words that are related to the word being disambiguated through semantic relations, such as hypernym, hyponym, holonym, meronym, troponym, and attribute of each word. This supplies a richer source of information and increases disambiguation accuracy. The adapted Lesk algorithm compares glosses between each pair of words in the window of context. These glosses are the ones associated with the synset, hypernym, hyponym, holonym,

Part of a subtitle file	*Extracted and tagged sentences*
1 00:00:25,600 –> 00:00:31,080 Human beings venture into the highest parts of our planet at their peril. 2 00:00:31,640 –> 00:00:34,480 Some might think that by climbing a great mountain 3 00:00:34,560 –> 00:00:36,320 they have somehow conquered it, 4 00:00:36,720 –> 00:00:39,800 but we can only be visitors here. 5 00:00:42,160 –> 00:00:46,240 This is a frozen alien world.	Human/JJ beings/NNS venture/NN into/IN the/DT highest/JJS parts/NNS of/IN our/PRP$ planet/NN at/IN their/PRP$ peril/NN. Some/DT might/MD think/VB that/IN by/IN climbing/VBG a/DT great/JJ mountain/NN they/PRP have/VBP somehow/RB conquered/VBN it/PRP, but/CC we/PRP can/MD only/RB be/VB visitors/NNS here/RB. This/DT is/VBZ a/DT frozen/JJ alien/JJ world/NN.

Fig. 2. Part of a subtitle file and its extracted sentences.

Part of a subtitle file	*Keywords assigned by TextRank*
Most of us would agree that a tiger is one of the world's most beautiful creatures. Sadly, in the wild, it's threatened with extinction. But, fortunately, it breeds very well in captivity, as this little cub proves. But is a tiger in a cage truly a tiger? I doubt it. To see the true essence and beauty of a tiger, you have to see it in the wild. This is the story of a tigress in the heart of India. Our tigress lives in Kanha National Park.	tiger, tigress, wild, national, kanha, captivity, breeds, lives, little

Fig. 3. Keywords of part of a subtitle file.

meronym, troponym, and attribute of each word. For example, the gloss of a synset of one word can be compared with the gloss of a hypernym of the other.

In our video categorization algorithm, WSD is essential for finding the WordNet domains of the words. Since we try to find the WordNet domains of keywords in the next step, we need to find the correct senses of these words. In our implementation, the correct sense of the keywords is assigned by using the adapted Lesk algorithm. For the keywords in Figure 3, the senses assigned by the adapted Lesk algorithm are given in Table 1.

By augmenting WordNet with domain labels, WordNet domains were created [34]. The synsets in WordNet have been annotated with at least one domain label by using a set of about 200 labels hierarchically organized. If there is no appropriate domain label for a synset, the label 'factotum' was assigned to it.

In the last step, the WordNet domains of the keywords are found. In finding the domains of a word, we should know the synset (gloss) of that word. Since we found the synsets of keywords in the WSD step, we made use of this information in finding the WordNet domains. The WordNet domains of the words are given in the last column of Table 1. Then, we calculated the occurrence score of each domain label (i.e. how many times a domain label appears in the keywords' domains) in a subtitle file and sorted them in a descending order.

We observed that video titles give important clues about categories of documentaries. For example, the category of the documentary 'Art of Spain' is 'Art'. As an extension to the approach of Katsiouli et al. [6], we decided to make use of the video title when categorizing the video which

Table 1
Senses of the keywords in Figure 3

Word	Pos Tag	Sense	Synset	WordNet Domains
tiger	Noun	1	tiger, Panthera tigris	animals, biology
tigress	Noun	0	tigress	animals
wild	Adjective	1	wild, untamed	factotum
national	Noun	0	national, subject	politics
kanha	Noun	–1	Not Found In WordNet Dictionary	—
captivity	Noun	0	captivity, imprisonment, incarceration, immurement	factotum
breeds	Verb	3	breed, multiply	factotum
lives	Verb	0	dwell, shack, reside, live, inhabit, people, populate, domicile, domiciliate	town_planning
little	Adjective	3	little, small	factotum

increased the performance of our algorithms. For this purpose, the WordNet domains are found for each word in the video title. Hence a list of WordNet domains which describes the video title is acquired. For example, the WordNet domains of the video title, 'Wildlife specials – tiger', are 'animals', 'biology' and 'factotum'.

Previously, we obtained WordNet domains of the video keywords and the occurrence scores of these domains. If one of these also exists in the video title domains, the occurrence score is increased by the ratio of one fourth. This ratio is determined experimentally. At the end of this step, we obtained the WordNet domains of a video with their occurrence scores.

3.2. Category label assignment method

Our first video categorization algorithm is category label assignment which uses mappings between categories and WordNet domains. In this algorithm, we took the approach of Katsiouli et al. [6] as a basis – some enhancements in implementing the steps were made and better results were obtained. In this video categorization algorithm, we find the WordNet domains related to the categories and a category label is assigned to the video by comparing the two. The overview of the algorithm is given in Figure 4.

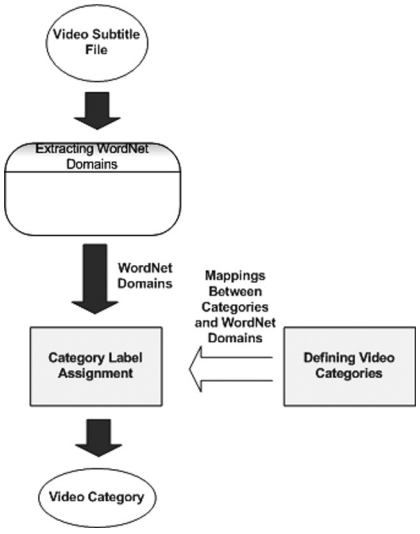

Fig. 4. Category label assignment.

Table 2
Category labels in the documentary collection and their corresponding WordNet domains

Category	Top rank WordNet domains
Geography	geography
Animals	animals, biology, entomology
Politics	politics, psychology
History	history, time_period
Religion	religion
Transportation	transport, commerce, enterprise
Accidents	transport, nautical
Sports	sport, play, swimming
War	military, history
Science	medicine, biology, mathematics
Music	music, linguistics, literature
Art	art, painting, graphic_arts
Technology	engineering, industry, computer_science
People	sociology, person

In order to assign a category label to a documentary video, a mapping is defined between the category labels and WordNet domains. First, the senses related to each category label were acquired from WordNet. The senses related with the category label through hypernym and hyponym relations and the WordNet domains corresponding to the senses of each category label were obtained. For each category, the occurrence scores of the derived domains were calculated and sorted in decreasing occurrence order. Table 2 shows the category labels and corresponding top ranked WordNet domains determined by Katsiouli et al. [6].

In the category label assignment step, a category label is assigned to the video. For a category label to be assigned, the sorted WordNet domains of the video were compared with the top rank domains of the categories. The algorithm compared the first domain of the video with the first domains of the categories.

- if the first domain of a category is equal to the first domain of the video, this category label is assigned to the video;
- if the first domain of more than one category is equal to the first domain of the video, the second domain of the corresponding sets are compared, and so on;
- if none of the category's first domains is equal to the first domain of the video, then the second domain of the video is compared to the first domain of the categories.

The algorithm continues as described above until a category label is assigned to the video. For example, when we consider the top rank WordNet domains for the categories in Table 3, if the sorted WordNet domains of a video are:

- 'animals, entomology, biology', then it is assigned to the 'Animals' category;
- 'transport, nautical, geography', then it is assigned to the 'Accidents' category;
- 'geography, animals', then it is assigned to the 'Animals' category.

At the text preprocessing step, while Katsiouli et al. [6] used Mark Hepple's POS tagger [38], we used the Stanford Log-linear Part-Of-Speech Tagger [36]. In the keyword extraction phase, they used one third of the number of words as the keyword count. In our system, we determined this number experimentally and observed that changing the number of keywords affected the system's classification accuracy (CA). Also in our implementation, we considered the effect of the video title since video titles give strong clues about the categories of documentary videos.

We implemented the approach of Katsiouli et al. [6] and evaluated with 130 documentary subtitles from National Geographic and the BBC. In this situation, we get 60% CA; after making changes to the algorithm this improved to 73.1% accuracy on the same experiment set.

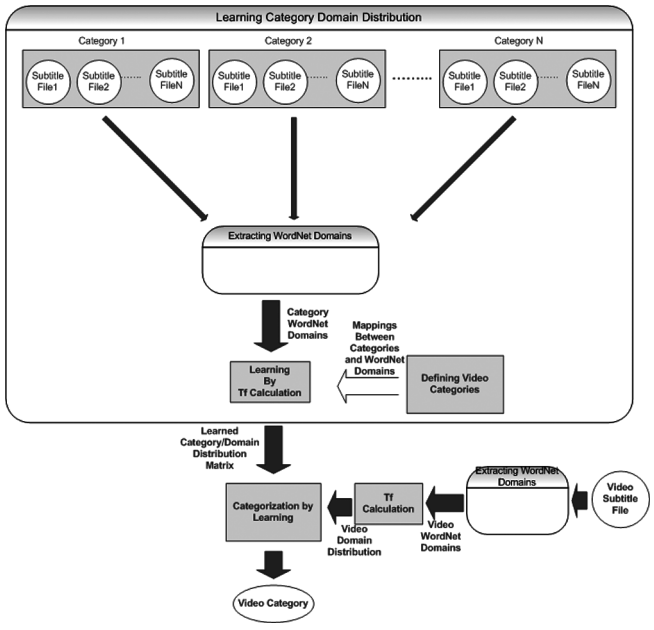

Fig. 5. Categorization by learning.

3.3. Categorization by learning method

Our second video categorization algorithm named as categorization by learning uses a learned category/domain distribution matrix. We propose a learning mechanism to assign a category label to videos. The preprocessing steps of the algorithm are the same as those used in the category label assignment method. This algorithm includes a learning phase. When a video is to be categorized, the domain distribution of the video is compared with the learned domain distribution of categories and the most similar category is assigned. The overview of the algorithm is given in Figure 5.

In the learning category domain distribution phase, documentaries with known categories are used as a training set. Our training set contains documentaries from all category labels. First of all, the documentary subtitles belonging to a specific category are processed using the extracting WordNet domains module. Hence, the domains and domain occurrence scores of the category are collected.

In order to determine the domain distribution of the category, we have used the term frequency (TF) weight of domains to determine the domain distribution of categories. The TF weight is computed for each category and domain pair. Hence a matrix showing the domain TF weights of all categories is obtained. Table 3 shows a sample part of the computed matrix representing the TF values of category domain pairs.

When we categorize a video, we compare the video's domain distribution with the domain distribution of categories and the category which has the most similar domain distribution with the video is selected. The subtitle of the video is processed in order to obtain the WordNet domains and the domain occurrence scores (TF values of the domains) of the video. Using the learned category/domain matrix, we try to find the category which has the most similar domain distribution to the video. For this purpose, we used the cosine similarity which is a measure of similarity between two vectors. For example, in order to categorize a documentary video named 'Everest', first we computed the domain distribution of the video. Table 4 shows the domain distribution of the documentary 'Everest'. Then, by using the learned matrix, we computed the similarities of categories. Table 5 shows the cosine similarities between the documentary 'Everest' and the categories. Since the 'Geography' category is most similar to the 'Everest' documentary, it is assigned as the documentary category.

Table 3
Sample part of the matrix representing the domain distribution of categories

	Geography	Animals	Politics
geography	0.0552444	0.032574	0.039201
animals	0.0302953	0.053447	0.009224
biology	0.0315682	0.041429	0.016141
entomology	0.0022912	0.000949	0
politics	0.0068737	0.008223	0.037663
psychology	0.0043279	0.007906	0.008455
history	0.0099287	0.008223	0.01691
time_period	0.0313136	0.023087	0.017294
religion	0.0089104	0.004744	0.021522
transport	0.0129837	0.012334	0.013451

Table 4
Domain distribution of the documentary 'Everest'

Domain	TF value
geography	0.036053131
animals	0.024667932
biology	0.032258065
entomology	0
politics	0.009487666
psychology	0.004743833
history	0.006641366
time_period	0.030360531
religion	0.011385199
transport	0.018975332
commerce	0.0028463
enterprise	0
nautical	0.003795066
sport	0.010436433
play	0.0056926
swimming	0
military	0.008538899
medicine	0.012333966
mathematics	0.0028463
music	0.0056926
linguistics	0.004743833
literature	0.004743833
art	0.003795066
painting	0
graphic_arts	0
engineering	0.000948767
industry	0

3.4. Experiments and evaluation

In order to evaluate the effectiveness of our categorization algorithms, we used documentaries from the BBC and National Geographic. The evaluation was performed using the CA metric, which reflects the proportion of the programme's correct assignments that agree with the original assignment.

For our first categorization algorithm (category label assignment), we conducted several experiments by changing some of the parameters. First of all, for keyword extraction, we changed the number of keywords selected and observed the results. As stated above, although Katsiouli et al. [6] selected a third of the words as keywords, we observed that this does not produce the best results.

Table 5
Cosine similarities between the documentary 'Everest' and the categories

Category	Cosine similarity
Geography	0.9590928
History	0.9452289
People	0.9376402
Animals	0.9267696
Science	0.9037822
Music	0.871545
Religion	0.825404
Politics	0.8154419
War	0.8115139
Art	0.7872766

Using the TextRank algorithm [10] all words are assigned a weight and selecting words above a certain weight could be an alternative for determining the number of keywords. Therefore, words above a certain weight are selected as keywords and the CA of the system is computed for changing weights. The diagram in Figure 6 shows the CA with changing keyword weights. For example, if we select the words with weight bigger than '5' as keywords, we get '50%' CA. As seen from Figure 6, we get the best results when using the weights between 0 and 0.4. Therefore using any weight between 0 and 0.4 does not change the CA, but selecting higher weights decreases the number of keywords. Using fewer keywords decreases the computation time. Therefore the upper bound value 'weight > 0.4' could be preferred to the others. So we used the experimentally determined value 'weight > 0.4'as the keyword selection parameter in our video categorization algorithm.

The effect of the video title when categorizing was subsequently considered. In the algorithm, we extracted the WordNet domains of a video and the occurrence scores of these domains. If one of these domains also existed in the video title domains, the occurrence score of increased by some ratio. The diagram in Figure 7 shows the effect of this ratio on the performance of the video categorization system. When the occurrence score of domains which also exist in the video title domains was increased by the ratio of 'one third' or 'one fourth', a CA of 75% was obtained. Selecting 'one third' or 'one fourth' does not make a significant difference in computation time, so any of them could be used in the video categorization algorithm – 'one fourth' was selected in our implementation.

In extracting the WordNet domains part of the categorization by learning algorithm, we used the parameters which obtained the best results in the first categorization experiment. Namely, keywords with 'weight > 0.4' in the TextRank algorithm and title effect by the ratio of 'one fourth'. To evaluate this algorithm, 65 documentaries were studied and categorized for learning purposes. An accuracy of 77% was achieved. It was noted that the performance of the system increases if the dataset used

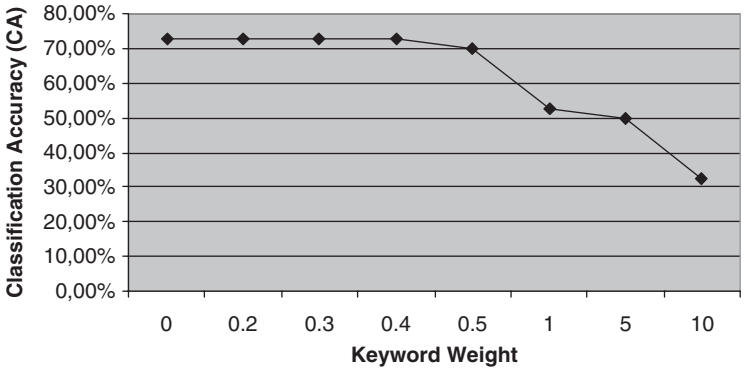

Fig. 6. Classification accuracy with keyword weights.

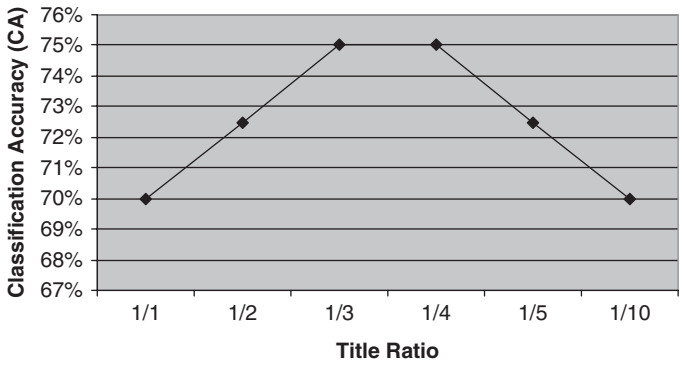

Fig. 7. Classification accuracy with title ratio.

for learning is enlarged. Also in our implementation, if more documentaries for learning could be used, better results would be maintained.

In order to see the performance of some of the well-known categorization algorithms on our subtitle dataset, we applied *K*-nearest neighbour (KNN) and SVM with polynomial kernel. We used half of the documents for training and half for the test set. KNN achieved an accuracy of 63%, and SVM achieved an accuracy of 70%. These results indicated that the usage of WordNet domains in our two categorization algorithms helps to increase the accuracy.

4. Video summarization

Video summarization algorithms present users with a condensed version of a video. In this paper, we used the subtitles of documentary videos to make summarizations. The summary sentences of the subtitle file were found by using text summarization techniques [10, 11], while the video segments corresponding to these summary sentences were extracted. By combining the video segments of summary sentences, we created a video summary. The overall approach is shown in Figure 8.

Subtitle files contain the text of the speech, the number and time of speech. In the text preprocessing step, the text in the subtitle file is extracted by striping the number and time of the speech, before being handed to the text summarization module. This module finds the summary sentences of the given text. Three algorithms were used to find the summary sentences: TextRank [10], Lexical Chain [11] and a combination of these two algorithms. After the summary sentences were found by one of these approaches, the text smoothing module applied some techniques to make summary sentences more understandable and smoother. The video summarization module used the summary sentences to create the video summary. The module found the start and end times of sentences from the video subtitle file. Then the video segments corresponding to the start and end times were subsequently extracted. By combining the extracted video segments, a video summary was generated.

4.1. Text summarization with the TextRank algorithm

The TextRank algorithm [10] extracts sentences for automatic summarization by identifying sentences that are more representative for the given text. To apply TextRank, we first built a graph and added a vertex to this graph for each sentence in the text. To determine the connection between vertices, we defined a 'similarity' relation between them, where 'similarity' is measured as a function of their content overlap. This relation can be thought of as a 'recommendation': a sentence mentioning certain concepts 'recommends' other sentences in the text that mention the same concepts and a connection is made The content overlap of two sentences is computed by the number of common tokens between them. To avoid promoting long sentences, the content overlap is divided by the length of each sentence.

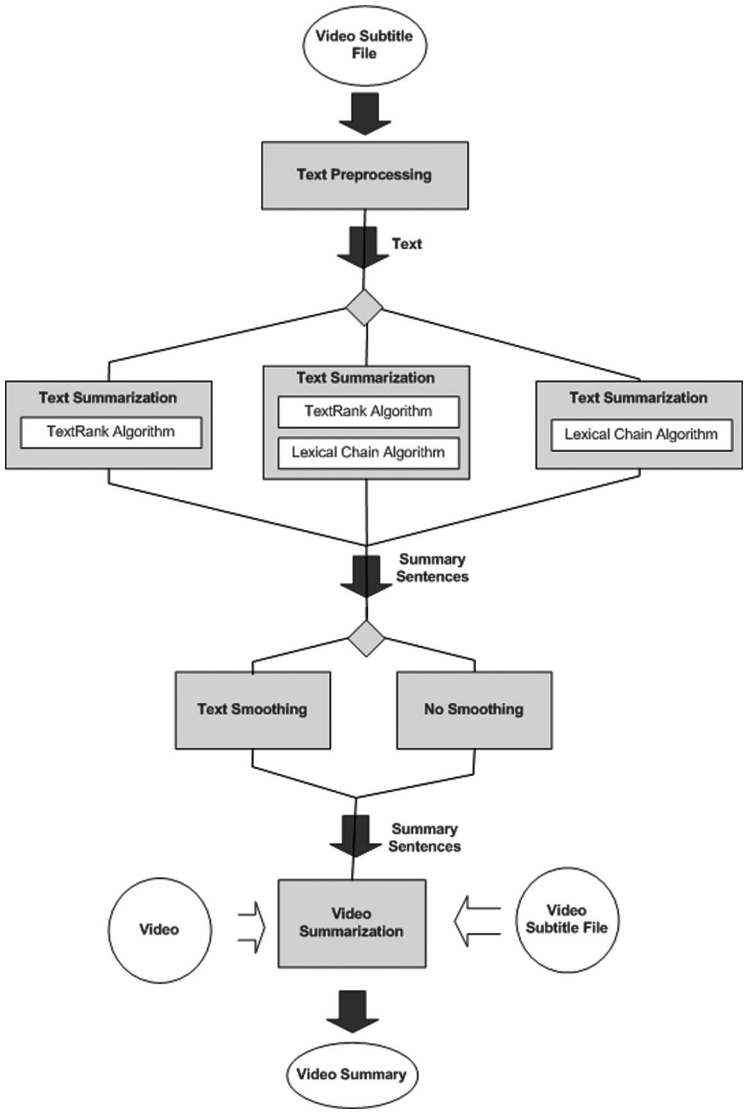

Fig. 8. Overall approach for video summarization.

A sentence composed of 'n' words is represented by $S_i = w_1,w_2,\ldots,w_n$, and two sentences S_i and S_j are given. Then the similarity of these sentences is defined formally as:

$$Similarity(S_i,S_j)=\frac{|\{w_k \mid w_k \in S_i \,\&\, w_k \in S_j\}|}{\log(|S_i|)+\log(|S_j|)}$$

The resulting graph is weighted since the edges have a similar weight. A weighted graph-based ranking algorithm is then used for deciding the importance of a vertex. Formally, the weighted score of a vertex V_i is defined as:

$$WS(V_i)=(1-d)+d*\sum_{v_j \in In(v_i)} \frac{w_{ji}}{\sum_{v_k \in Out(V_j)} w_{jk}} WS(V_j)$$

Here, $In(V_i)$ is the set of vertices that point to V_i and $Out(V_i)$ is the set of vertices that V_i points to. The weight of the edge between the vertices V_i and V_j is w_{ji}, and d is the damping factor that can be set between 0 and 1. The value of d is usually set to 0.85 and this value is also used in our implementation. After the ranking algorithm, sentences are sorted using their score and top ranked sentences are selected as the summary sentences.

4.2. Text summarization with the lexical chain algorithm

In [11], automated text summarization is done by identifying the significant sentences of text. The lexical cohesion structure of the text is exploited to determine the importance of sentences. Lexical chains can be used to analyse the lexical cohesion structure in the text. In the proposed algorithm, first, the lexical chains in the text are constructed. The lexical chaining algorithm is an implementation of Galley et al.'s algorithm [39] with some small changes. Topics are then roughly detected from lexical chains and the text is segmented with respect to the topics. It is assumed that the first sentence of a segment is a general description of the topic, so the first sentence of the segment is selected as the summary sentence.

4.3. Text summarization with a combination of algorithms

We proposed a new summarization approach by combining the two summarization algorithms, the TextRank algorithm [10] and the lexical chain algorithm [11]. Once the summary sentences of a text were found we determined the common sentences of the two summaries and selected these to be included in the main summary. Both algorithms sorted the summary sentences with respect to their importance. After selecting the common sentences, the most important sentences of the two algorithms up to the length of the desired summary were extracted. An overview of the summarization, with the combination of the algorithms, is given in Figure 9.

4.4. Text smoothing

In order to improve the understandability and completeness of the summary, some smoothing operations were carried out after the text summarization phase. It is observed that some of the selected

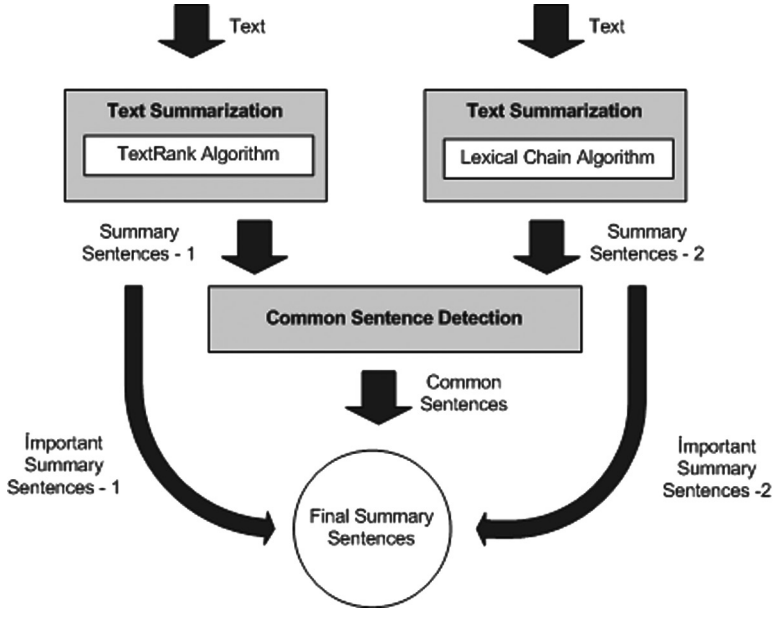

Fig. 9. Overview of the text summarization by combination of algorithms.

sentences start with a pronoun and if they are not included in the previous sentences in the summary these pronouns may be confusing.

In order to handle this problem, if a sentence starts with a pronoun, the preceding sentence is also included in the summary. If the preceding sentence also starts with a pronoun, its preceding sentence is also added to the summary sentence list. The backward processing of the sentences contains only two steps. We observed that, if a sentence starts with a pronoun, including just the preceding sentence solves the problem in most cases and the summary becomes more understandable.

4.5. Clip generation

Our video summarization approach is based on the summary sentences found by the text summarization algorithms. After finding the summary sentences, the start and end times of these sentences are found from video subtitle file. For each summary sentence, the video segment corresponding to the sentence is extracted from the video by using the start and end time of the sentence. Then, by combining the extracted video parts, a video summary is created. A screenshot of our video summarization system is presented in Figure 10. The system lets the user select the summarization algorithm from the summary method group box. The user can select the algorithms 'TextRank', 'LexicalChain' or 'Mixed'. The user can also select the options 'Normal' or 'Smooth'. The former indicates that the summarization system will not use text smoothing after text summarization.

4.6. Experiments and evaluation

The evaluation of video summaries is difficult because they are so subjective. Different people will compose different summaries for the same video. The evaluation of video summaries could be conducted by requesting people watch the summary and asking them several questions about the video. However, in our summarization system, since we used text summarization algorithms, we prefered to evaluate the text summarization algorithms only. We believe that the success of the text summarization directly determines the success of video summarization in our system. For the evaluation of text summarization, we use ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [40], which makes evaluation by comparing the system generated output summaries to model summaries written by humans.

Fig. 10. Video summarization system screenshot.

Table 6
ROUGE scores of algorithms in a video summarization system

	ROUGE-1	ROUGE-L	ROUGE-W
TextRank	0,33877	0,33608	0,13512
TextRank_Smooth	0,34453	0,34184	0,13686
LexicalChain	0,24835	0,24600	0,10413
LexicalChain_Smooth	0,25211	0,24976	0,10529
Mix	0,34375	0,34140	0,13934
Mix_Smooth	0,34950	0,34716	0,14108

ROUGE is the most popular summarization evaluation methodology. All of the ROUGE metrics aim to find the percentage of overlap between the system output and the model summaries. ROUGE calculates the ROUGE-N score (calculated using N-grams), ROUGE-L score (calculated using longest common subsequences) and ROUGE-W score (calculated using weighted longest common subsequences).

In our video summarization system, we used six approaches for finding the summary of the subtitle text of a video:

- the TextRank algorithm;
- the TextRank algorithm and smoothing the result;
- the LexicalChain algorithm;
- the LexicalChain algorithm and smoothing the result;
- a mix of the TextRank and LexicalChain algorithms;
- a mix of the TextRank and LexicalChain algorithms and smoothing the result.

All six approaches were studied using the BBC documentaries. Students were asked to compose summaries of the selected documentaries by selecting the 20 most important sentences from the subtitles. The same documentaries were also summarized by our video summarization system, using the algorithms mentioned above, which generated summaries composed of 20 sentences. We calculated ROUGE scores in order to compare the system outputs with human summaries. While calculating ROUGE scores, we applied Porter's Stemmer and stop word list on the input. ROUGE scores of the algorithms in our video summarization system can be seen in Table 6. We observed that smoothing improves the performance of all algorithms. When the TextRank algorithm is used, better results were obtained than when the LexicalChain algorithm was implemented. The best results were obtained by using the mixed TextRank and LexicalChain algorithms to find the summary sentences and smoothing the results. Our best ROUGE scores were comparable with the ROUGE scores of the state of the art systems in the literature.

5. Conclusions

This paper presented a system which performs automatic categorization and summarization of documentary videos with subtitles. We wanted to handle these problems together because their outputs support each other. Presenting both the category and the semantic summary of a video would give viewers quick and satisfactory information about that content.

The automatic video categorization was performed by two categorization methods, category label assignment and categorization by learning. The CA of the former was evaluated on documentary videos and promising results were obtained. The second method used a limited number of videos for learning. In future work, we want to improve this by using more videos. It is known that using more data for learning increases the performance of the system and gives better results.

We performed video summarization by using video subtitles and employing text summarization methods. Two text summarization algorithms [10, 27] were used and their results were applied to

the video summarization domain. In this work, we took advantage of the characteristics of the documentary videos where the speech and display of the video have a strong correlation.

Video summary is produced by extracting the video parts corresponding to the summary sentences. This extraction could be improved by employing a shot identification mechanism. An extracted video part could be extended by finding the start and end of the residing shot. In this way, the video parts could show a more complete presentation. In the evaluation of video summaries, the programme summaries were compared with human generated summaries and the ROUGE score recorded. In future work, we want to perform the evaluation by using the video summaries alongside the text summaries. Video summaries could be watched by viewers who could then evaluate the results.

Both algorithms are currently used in English, but it is possible to convert them into different languages. The language dependency of the algorithms is caused by the WordNet and the natural language processing (NLP) tools such as POS tagger. If the WordNet and the required NLP tools are available for other languages, our video categorization and summarization algorithms can be used for videos with other language subtitles.

Acknowledgements

This work is partially supported by the Scientific and Technical Council of Turkey Grant TUBITAK EEEAG-107E234, and the Scientific and Technical Council of Turkey Grant TUBITAK EEEAG-107E151.

References

[1] D. Brezeale, D.J. Cook, Automatic video classification: a survey of the literature, *IEEE Transactions Systems, Man and Cybernetics Part C: Applications and Reviews* 38(3) (2008) 416–430.

[2] S. Teufel and M. Moens, Sentence extraction as a classification task, *Proceedings of ACL/EACL 97 WS* (Madrid, Spain, 1997).

[3] X. Yuan, W. Lai, T. Mei, X.S. Hua, X.Q. Wu and S. Li, Automatic video genre categorization using hierarchical SVM, *Proceedings of IEEE International Conference on Image Processing* (2006) 2905–2908.

[4] W. Zhu, C. Toklu and S.P. Liou, Automatic news video segmentation and categorization based on closed-captioned text, *ISIS Technical Report Series* 20 (2001).

[5] F.N. Bezerra and E. Lima, Low cost soccer video summaries based on visual rhythm, *Proceedings of the 14th Annual ACM International Conference on Multimedia* (Santa Barbara, CA, 23–27 October 2006) 71–77.

[6] P. Katsiouli, V. Tsetsos and S. Hadjiefthymiades, Semantic video classification based on subtitles domain terminologies, *Proceedings of SAMT Workshop on Knowledge Acquisition from Multimedia Content (KAMC)* (Genoa, Italy, 2007).

[7] C. DeMenthon, V. Kobla and D. Doermann, Video summarization by curve simplification, *Proceedings of ACM Multimedia* (1998) 211–218.

[8] B. Barbieri, N. Dimitrova and L. Agnihotri, Movie-in-a-minute: automatically generated video previews, *Proceedings of IEEE Pacific Rim Conference on Multimedia* (9–18 February 2004).

[9] K. Fujimura, K. Honda and K. Uehara, Automatic video summarization by using color and utterance information, *Proceedings of IEEE ICME* (2002) 49–52.

[10] R. Mihalcea and P. Tarau, TextRank – bringing order into texts, *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Barcelona, Spain, 2004).

[11] G. Ercan and I. Cicekli, Lexical cohesion based topic modeling for summarization, *Proceedings of the CICLing* (2008) 582–592.

[12] P. Wang, R. Cai and S.Q. Yang, A hybrid approach to news video classification multimodal features, *Proceedings of the Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing and the Fourth Pacific Rim Conference on Multimedia* (2003) 787–791.

[13] W. Qi, L. Gu, H. Jiang, X.R. Chen and H.J. Zhang, Integrating visual, audio and text analysis for news video, *Proceedings of 7th IEEE International Conference Image Processing* (2000) 520–523.

[14] N. Benjamas, N. Cooharojananone and C. Jaruskulchai, Flashlight and player detection in fighting sport for video summarization, *Proceedings of the IEEE International Symposium on Communications and Information Technology* (Beijing, China, 12–14 October 2005) 441–444.

[15] A. Money and H. Agius, Video summarisation: a conceptual framework and survey of the state of the art, *Journal of Visual Communication and Image Representation* 19(2) (2008) 121–143.
[16] A. Ekin, M. Tekalp and R. Mehrotra, Automatic soccer video analysis and summarization, *IEEE Transactions on Image Processing* 12(7) (2003) 796–807.
[17] G. Ciocca and R. Schettini, Dynamic storyboards for video content summarization, *Proceedings of the 8th ACM SIGMM International Workshop on Multimedia Information Retrieval* (Santa Barbara, CA, 26–27 October 2006).
[18] Z. Cernekova, I. Pitas and C. Nikou, Information theory-based shot cut/fade detection and video summarization, *IEEE Transactions on Circuits and Systems for Video Technology* 16(1) (2006) 82–91.
[19] A. Girgensohn, A fast layout algorithm for visual video summaries, *Proceedings of the IEEE International Conference on Multimedia and Expo* (Baltimore, MD, 6–9 July 2003) 77–80.
[20] B. Ngo, Y. Ma and H. Zhang, Video summarization and scene detection by graph modeling, *IEEE Transactions on Circuits and Systems for Video Technology* 15(2) (2005) 296–305.
[21] Y. Rui, A. Gupta and A. Acero, Automatically extracting highlights for TV baseball programs, *Proceedings of the 8th ACM International Conference on Multimedia* (Los Angeles, CA, 30 October 2000) 105–115.
[22] M. Pickering, L. Wong and S. Ruger, ANSES: summarisation of News Video, *Proceedings of CIVR-2003* (University of Illinois, IL, 24–25 July 2003).
[23] T. Tsoneva, M. Barbieri and H. Weda, Automated summarization of narrative video on a semantic level, *Proceedings of the International Conference on Semantic Computing* (17–19 September 2007) 169–176.
[24] F.N. Bezerra and E. Lima, Low cost soccer video summaries based on visual rhythm, *Proceedings of the 14th Annual ACM International Conference on Multimedia* (Santa Barbara, CA, 23–27 October 2006) 71–77.
[25] J. Kupiec, J.O. Pedersen and F. Chen, A trainable document summarizer, *Proceedings of SIGIR 1995* (ACM Press, New York, 1995) 68–73.
[26] D.R. Radev, H. Jing and M. Budzikowska, Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies, *Proceedings of ANLP/NAACL00-WS*, (Seattle, WA, 2000).
[27] G. Erkan and D.R. Radev, Lexrank: graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* 22 (2004) 457–479.
[28] L. Page, S. Brin, R. Motwani and T. Winograd, The pagerank citation ranking: bringing order to the web, Technical Report, *Stanford Digital Library Technologies Project*, 1998.
[29] R. Barzilay and M. Elhadad, Using lexical chains for text summarization. In: I. Mani and M.T. Maybury (eds), *Advances in Automatic Text Summarization* (The MIT Press, Cambridge, MA, 1999) 111–121.
[30] M. Brunn, Y. Chali and C.J. Pinchak, Text summarization using lexical chains, *Proceedings of the Document Understanding Conference (DUC01)* (New Orleans, LA, 2001).
[31] W.P. Doran, N. Stokes, J. Carthy and J. Dunnion, Assessing the impact of lexical chain scoring methods and sentence extraction schemes on summarization, *Proceedings of CICLing* (2004) 627–635.
[32] G.H. Silber and K. McCoy, Efficient text summarization using lexical chains, *Proceedings of the ACM Conference on Intelligent User Interfaces (IUI'2000)*, 9–12 January 2000.
[33] Y. Chali and M. Kolla, University of Lethridge summarizer at DUC04, *Proceedings of DUC04* (Boston, July 2004).
[34] L. Bentivogli, P. Forner, B. Magnini and E. Pianta, Revising WordNet Domains hierarchy: semantics, coverage, and balancing, *Proceedings of COLING Workshop on Multilingual Linguistic Resources* (Geneva, 2004) 101–108.
[35] S. Banerjee and T. Pedersen, An adapted Lesk algorithm for word sense disambiguation using WordNet, *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-02)* (Mexico City, 2002).
[36] *Stanford Log-linear Part-Of-Speech Tagger*, Available at: http://nlp.stanford.edu/software/tagger.shtml
[37] *Penn Treebank*, Available at: www.cis.upenn.edu/~treebank/
[38] M. Hepple, Independence and commitment: assumptions for rapid training and execution of rule-based part-of-speech taggers, *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* (Hong Kong, 2000).
[39] M. Galley and K. McKeown, Improving word sense disambiguation in lexical chaining, *Proceedings of the 18th International Joint Conference on Artificial Intelligence* (2003) 1486–1488.
[40] C.Y. Lin and E.H. Hovy, Automatic evaluation of summaries using n-gram co-occurrence statistics, *Proceedings of HLT-NAACL-2003* (Edmenton, Canada, 2003).