Neural Network-Based Movie Dialogue Detection

Margarita Kotti, Emmanouil Benetos, Constantine Kotropoulos
Department of Informatics, Aristotle Univ. of Thessaloniki
Box 451, Thessaloniki 541 24, Greece
E-mail: {mkotti,empeneto,costas}@aiia.csd.auth.gr

Abstract

A novel framework for dialogue detection in movies, using indicator functions, is investigated. An indicator function determines if an actor is present at a particular time instant. The cross-correlation function of a pair of indicator functions and its related cross-power spectral density are applied as inputs to neural networks. Several types of neural networks are employed to test the feasibility of the proposed framework, such as perceptrons, radial-basis function networks, and support vector machines. Experiments are conducted on indicator functions extracted from 6 different movies that correspond to a total of 41 dialogue instances and 20 non-dialogue instances. High accuracy detection is achieved on average, ranging between 84.780%±5.499% and 94.740%±5.263%, with a mean value of 88.990%±2.967%.

1. Introduction

Nowadays movie archives have become a common place. It is estimated that over 9.000 hours of video are released every year. Due to their large size efficient handling, searching, indexing, browsing, summarization and retrieval is rather cumbersome. Consequently, movie content analysis is a necessity as can one realize by the creation of the MPEG-7 standard (formerly known as Multimedia Content Description Interface) [21].

It is true that up-to-date approaches for automatic movie analysis concentrate mainly on the visual channel and tend to neglect the corresponding audio channel. However, as it is concluded in this paper, the audio channel can significantly contribute to movie content analysis. Accordingly, combined audio and video analysis are expected to obtain improved results than video channel or audio channel analysis individually. Related topics to dialog detection are face detection and tracking, speaker turn detection [10], and speaker tracking [14].

Numerous methods for dialogue detection can be found in the bibliography. For example, low-level audio and visual features achieved a maximum classification accuracy of 96% in [1]. A maximum recall equal to 0.880 is attained by fusion of video and audio information in [9]. Additionally, emotional stages as means for video segmenting have been employed in [19].

Movie dialog detection follows specific rules, since movie making is a kind of art and it has its own grammar [3]. There are various definitions for dialogue scenes. A dialogue scene can be described as a set of consecutive shots which contain people conversations [2]. Another definition suggests that the elements of a dialog scene are: the people, the conversation, and the location where the dialog is taking place [5]. Lehane suggests an additional definition of dialog scenes, claiming that in a 2-person dialog there is usually an A-B-A-B structure of camera angles[13]. However, this is not the only case, since the person who speaks at a given time is not always the one displayed. For example, shots of another participant reactions are frequently inserted. In addition, the shot of the speaker may not include his face, but the back view of his head. Various shots may be entered in the dialog scene, such as other persons or objects. Evidently, these shots add to the complexity of the dialog detection problem, due to their nondeterministic nature. In this paper, we discriminate the audio types as: (i) Clean dialogue: Dialogues with low-level audio background, (ii) Dialogue with background: Dialogue in the presence of a noisy background or music, (iii) Clean monologue: Monologues with low-level audio background, (iv) Monologue with background: Monologue in the presence of a noisy background or music, (v) Other: Anything else except the above.

In this paper, we apply *indicator functions* as means of *audio-assisted* dialogue detection. Ground truth indicator functions, extracted by human observers, are employed. This experimental protocol corresponds to the ideal situation where the indicator functions are *error free*. In practice, indicator functions can be obtained, for example, by speaker turn detection followed by speaker clustering, which is the case of speaker diarization. As inputs to neu-

ral networks, the cross-correlation values of a pair of indicator functions and the magnitude of the corresponding cross-power spectral density are utilized. Neural networks employed are: perceptrons and radial-basis function networks. Support vector machines are used as well. Experiments are conducted using audio scenes extracted from six different movies, as can be seen in Table 1. In total, 25 dialog and 17 non-dialog scenes are employed. Dialogs last from 20 sec to 123 sec. In this paper we concluded that an adequate duration to identify a dialogue scene is 25 sec. Applying a properly chosen time window to the scenes leads to a total of 41 dialog instances and another 20 non-dialog instances. A high dialog detection accuracy is achieved, ranging between 84.780%±5.499% and 94.740%±5.263%, with a maximum F_1 measure of 0.958.

The remainder of the paper is as follows. In Section 2, the notion of indicator functions is introduced in the framework of dialog detection. In addition, cross-correlation and cross-power spectral density are described as features for dialog detection. The dataset created is detailed in Section 3 and the applied figures of merit in Section 4. In Section 5, the experimental results using artificial neural networks are described and their performance is discussed. Finally, conclusions are drawn in Section 6.

2. Dialogue detection

2.1 Indicator Functions

Let us consider that for an audio recoding of N samples it is known when a particular actor (i.e. speaker) performs. This information can be quantified by the indicator function of say actor A, $I_A(n)$, defined as:

$$I_A(n) = \begin{cases} 1, & \text{actor } A \text{ is present at sample } n \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

In this paper, 2-actors dialogues are assumed for simplicity reasons but without loss of generality. For the case of two actors A and B, the corresponding indicator functions are $I_A(n)$ and $I_B(n)$, respectively. In every-day dialogs, the first actor rarely stops at time n and the second actor starts at time $n+1$. As a results, audio frames corresponding to both are possibly present. Additionally, short silence periods should be tolerated. Moreover, dialog detection should not be inhibited by background music, environmental noise etc. Monologues, music soundtrack, songs, street noise, or instances where the first actor is talking and the second one is just making exclamations are some examples of non-dialogue scenes. In this paper, optimal, error-free (i.e. ground-truth) indicator functions are employed. Ongoing research employs indicator functions that are determined by diarization. This way, no human interaction is needed.

A characteristic plot of indicator functions in a dialogue case is shown in Figure 1. In Figure 2, a typical example of a non-dialog (i.e. a monologue) is depicted, where $I_B(n)$ corresponds to short exclamations of the second actor.

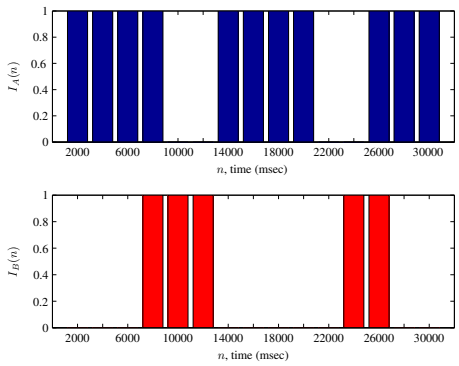

Figure 1. Indicator functions of two actors in a dialog scene.

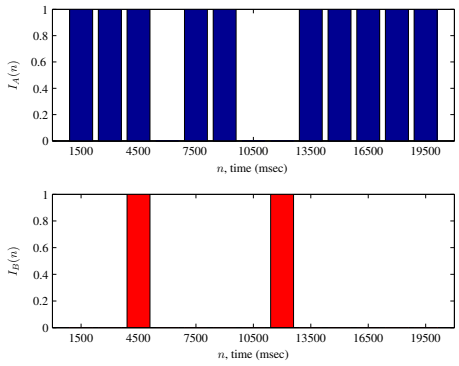

Figure 2. Indicator functions of two actors in a non-dialog scene (i.e. monologue).

2.2 Cross-correlation and Cross-power Spectral Density

The cross-correlation is a widely used measure of similarity between two signals [17]. It is commonly applied to find the linear relationship of two signals. In the case of a pair of indicator functions, the cross-correlation is:

$$c_{AB}(d) = \begin{cases} \frac{1}{N}\sum_{n=1}^{N-d} I_A(n+d)I_B(n), \\ \qquad \text{when } 0 \le d \le N-1 \\ c_{BA}(-d), \\ \qquad \text{when } -(N-1) \le d \le 0 \end{cases} \quad (2)$$

where d is the time-lag. The presence of a dialog is deducted by significantly large values of the cross-correlation. A representative cross-correlation function for a dialog scene is depicted in Figure 3. It is for the same audio stream, whose indicator function is shown in Figure 1.

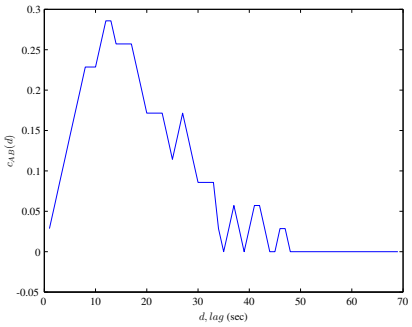

Figure 3. Cross-correlation of the indicator functions of two actors in a dialog case.

The second function examined is the cross-power spectral density i.e. the discrete-time Fourier transform of the cross-correlation [17]. The cross-power spectral density is defined as:

$$\phi_{AB}(f) = \sum_{d=-(N-1)}^{N-1} c_{AB}(d)\ \exp\left(-j2\pi\ f\ d\right) \quad (3)$$

where $f \in [-0.5, 0.5]$ is the frequency in cycles per sampling interval. For negative frequencies, it holds $\phi_{AB}(-f) = \phi^*_{AB}(f)$, with * standing for the complex conjugation operator. The magnitude of the cross-power spectral density is used, which is the common case in audio processing. If the area under $|\phi_{AB}(f)|$ is considerably large, a dialogue scene is presumed to be present, whereas for a non-dialog case $|\phi_{AB}(f)|$ value is most likely to be small. In Figure 4 the magnitude of the cross-power spectral density is demonstrated. It is for the same audio stream, whose cross-correlation is depicted in Figure 3

In preliminary experiments on dialog detection, only two values were used: the cross-correlation value at zero lag $c_{AB}(0)$ and the cross-spectrum energy in the frequency band [0.065, 0.25] [11]. The frequency is measured in cycles per sampling interval and the values 0.065 and 0.25

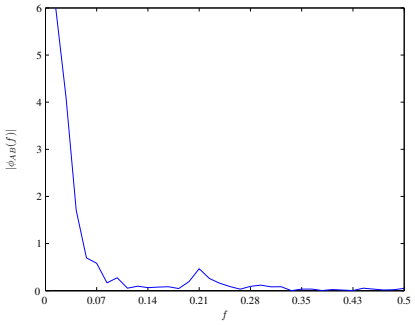

Figure 4. Magnitude of the cross-power spectral density for two actors in a dialog case.

were experimentally chosen. Subsequently, to alleviate this fixed choice, in this work we resort to the sequences of the cross-correlation and the cross-power spectral density. A pair of thresholds were compared against the aforementioned values. The thresholds were determined after training aiming to detect dialogs. In this paper, a more generic approach is applied. The cross-correlation sequence is evaluated over properly chosen time-windows. That is also the case for magnitude of its Discrete Fourier Transform, i.e. the uniform frequency sampling of the cross-power spectral density.

3 Dataset

Audio scenes from 6 movies are extracted to create the dataset, as presented in Table 1. The specific movies are chosen for various reasons. Firstly, they are popular and consequently easily accessible. Secondly, they cover a wide area of movie genres. For example, Analyze That is a comedy, Platoon is an action, and Cold Mountain is a drama. Finally, they have already been widely employed for movie analysis experiments. From the aforementioned movies a total of 42 scenes is extracted. This is the largest movie set utilized in audio-assisted dialogue detection, to the best of the authors' knowledge. Dialogue scenes are 25, while the remaining 17 are non-dialogue scenes, as can be seen in Table 1. The audio track is digitized in PCM at a sampling rate of 48 kHz and each sample is quantized in 16 bit two-channel. All 42 scenes have a duration of 34 min and 43 sec.

4 Figures of Merit

Commonly used figures of merit for dialog detection are summarized in this Section to enable a comparable performance assessment with other similar works. Let $hits_d$ be

Table 1. The six movies used to create the dataset.

Movie name	Dialog scenes	Non-dialog scenes	Total scenes
Analyze That	4	2	6
Cold Mountain	5	1	6
Jackie Brown	3	3	6
Lord of the Rings I	5	3	8
Platoon	4	2	6
Secret Window	4	6	10
Total	25	17	42

the correctly classified dialog instances and $hits_{nd}$ the correctly classified non-dialog instances. Dialog instances that are not classified correctly are $misses$ and $false\ alarms$ are the non-dialog instances classified as dialog ones. Obviously, the total number of dialog instances is equal to the sum of $hits_d$ plus $misses$.

Two sets of figures of merit are employed. The percentage of correctly classified instances (CCI) and the root mean squared error ($RMSE$) are contained in the first one. CCI is given by:

$$CCI = \frac{hits_d + hits_{nd}}{hits_d + hits_{nd} + misses + false\ alarms} \cdot 100\%. \quad (4)$$

$RMSE$ for the 2-class classification, is:

$$RMSE = \sqrt{\frac{misses + false\ alarms}{hits_d + hits_{nd} + misses + false\ alarms}} \cdot 100\%. \quad (5)$$

Another commonly used triplet is precision (PRC), recall (RCL), and F_1 measure. For the dialog instances, they are defined as:

$$PRC = \frac{hits_d}{hits_d + false\ alarms} \quad (6)$$

$$RCL = \frac{hits_d}{hits_d + misses}. \quad (7)$$

F_1 measure admits a value between 0 and 1. It is defined as:

$$F_1 = \frac{2\ PRC \cdot\ RCL}{PRC + RCL}. \quad (8)$$

The higher its value is, the better performance is obtained.

5. Experimental results and performance evaluation

To fix the number of inputs of the neural networks, a running time-window is applied to each audio scene. The running time-windows exhibit no overlap. The mean actor utterance duration has been found to be about 5 sec. 4 actor changes are presumed to occur within the running time-window employed in our analysis, on average. Accordingly, an A-B-A-B-A structure is expected for a dialogue to occur. Similar assumptions were also invoked in [13, 20]. As a result, an appropriate time-window should have a duration of $5 \times (4 + 1)$ = 25 sec. After applying the 25 sec running time-window to the 42 audio scenes, 61 instances are extracted. 41 dialog instances and 20 non-dialog distances are obtained. For a 25 sec window and a sampling frequency of 100 Hz, 49 samples of $c_{AB}(d)$ and another 49 samples of $|\phi_{AB}(f)|$ are computed. The aforementioned 98 samples, are fed as input to the neural networks, while the label stating whether the instance is a dialog or not, is exploited only during training. The experiments are repeated 7 times and the averaged values of figures of merit are recorded. For comparison reasons, two commonly used splits between the training and test sets are utilized: the 70%/30% and 50%/50% ratios. In the first case, the range of the error equals 10%, while in the second case equals 5%.

5.1 Perceptrons

Two perceptron network variants are discussed: the multilayer perceptron (MLP) network and the voted perceptron (VP) network.

MLPs are feed-forward networks, consisting of multiple layers of computational units. There are three layers in the case under consideration: the input layer consists of 98 input nodes (i.e. 49 for $c_{AB}(d)$ and another 49 for $|\phi_{AB}(f)|$), the hidden layer, consists of 51 nodes, and the output layer. The number of hidden nodes was determined experimentally. The sigmoid function is applied as the activation function.

Here, two MLP categories are studied. For the first category the learning technique is the back-propagation algorithm. Back-propagation MLPs tend to overfit the training data, particularly when the size of the training set is restrained. Also, computation speed and convergence problems rise. The optimization problem, with respect to the computational cost can be solved by utilizing the fast artificial neural network library (FANN) [15]. 7-repeats average dialogue detection efficiency results using back-propagation MLPs with 70%/30% and 50%/50% training/test set splits are depicted in Table 2.

The second category of MLP networks employs particle swarm optimization (PSO) as a replacement of the back-propagation algorithm. PSO is an algorithm inspired by the social behavior of bird flocks and fish schools [6]. Every candidate solution is called a particle and has a current position and a velocity. The currently optimum particles are followed through the problem hyperspace by the remain-

Table 2. Averaged figures of merit for dialog detection using back-propagation MLPs for 70%/30% and 50%/50% training/test set splits after 7 repetitions.

Figures of merit	70%/30%	50%/50%
CCI (mean)	90.970%	86.170%
CCI (st. dev.)	3.976%	5.172%
$RMSE$	0.259	0.326
PRC	0.978	0.948
RCL	0.892	0.843
F_1	0.931	0.890

ing particles. Although PSO and genetic algorithms have many similarities, no evolution operators, such as crossover and mutation are present in the PSO case. PSO compared to back-propagation MLP networks deals more efficiently with the problem of computation time, yields improves results, do not overfit the data, and approximates better a nonlinear function thus exhibiting a better global convergence. For the case under consideration, a 3-layered feed-forward network is utilized. Trelea type-II PSO is employed for learning [18]. 7-repeats averaged results on dialog detection for the 3-layered PSO-trained MLP network are depicted In Table 3.

Table 3. Averaged figures of merit for dialog detection using a 3-layered PSO-trained MLP feed-forward network for 70%/30% and 50%/50% training/test set splits after 7 repetitions.

Figures of merit	70%/30%	50%/50%
CCI (mean)	88.880%	91.430%
CCI (st. dev.)	4.535%	4.239%
$RMSE$	0.326	0.283
PRC	0.895	0.900
RCL	0.982	0.987
F_1	0.934	0.941

The second perceptron network variant applied is VP. VP is easy to implement and computation time economical. The leave-one-out method is used by VP. In VP, the algorithm takes advantage of data that are linearly separable with large margins [7]. Its philosophy is based on the assumption that data are more likely to be linearly separable into higher dimension spaces. For the marginal case of one epoch, VP is equivalent to MLP. 7-repeats averaged dialogue detection results for the two splits are enlisted in Table 4.

Table 4. Averaged figures of merit for dialog detection using VPs for 70%/30% and 50%/50% training/test set splits after 7 repetitions.

Figures of merit	70%/30%	50%/50%
CCI (mean)	88.720%	86.630%
CCI (st. dev.)	6.393%	4.337%
$RMSE$	0.305	0.360
PRC	0.864	0.849
RCL	0.998	0.979
F_1	0.920	0.908

5.2 Radial Basis Functions

Radial basis functions (RBFs) can replace the sigmoidal hidden layer activation function in MLPs. Compared to MLP networks, RBF networks do not face the problem of local minima because the linear mapping from the hidden layer to the output layer is adjusted in the learning process. In this paper, a normalized Gaussian RBF network is utilized. Basis functions is based on the k-means clustering algorithm. The logistic regression model is employed for learning [8]. The data of each cluster is fit to symmetric multivariate Gaussians. Standardization takes place, leading all features to have zero mean and unit variance. Averaged dialog detection results for the two splits using the RBF network are summarized in Table 5.

Table 5. Averaged figures of merit for dialog detection using RBF networks for 70%/30% and 50%/50% training/test set splits after 7 repetitions.

Figures of merit	70%/30%	50%/50%
CCI (mean)	87.210%	84.780%
CCI (st. dev.)	5.135%	5.499%
$RMSE$	0.318	0.357
PRC	0.908	0.923
RCL	0.913	0.855
F_1	0.906	0.885

5.3 Support Vector Machines

Support vector machines (SVMs) are supervised learning methods that can be applied either to classification or regression. They are related to RBF networks, although SVMs in order to avoid over-fitting, try to detect the maximum-margin hyperplane. Here, a second order polynomial kernel is applied. The sequential minimal optimization algorithm is used for training the support vector classifier [4, 16]. The penalty parameter C is set equal to 1. Smaller values that 1 are found to deteriorated results, whereas larger values that 1 yield the same performance. 7-repeats averaged dialogue detection experimental results using SVMs with 70%/30% and 50%/50% training/test set splits are detailed in Table 6.

Table 6. Averaged figures of merit for dialog detection using the SVM classifier for 70%/30% and 50%/50% training/test set splits after 7 repetitions.

Figures of merit	70%/30%	50%/50%
CCI (mean)	94.740%	90.320%
CCI (st. dev.)	5.263%	9.677%
$RMSE$	0.2929	0.311
PRC	1.000	0.955
RCL	0.917	0.913
F_1	0.958	0.933

5.4 Performance Evaluation

Average detection accuracy demonstrates a mean value of $88.990\% \pm 2.967\%$. The SVM detection accuracy for the 70%-30% split is the highest achieved, with F_1 measure equal to 0.958. This may be attributed to the fact that SVMs are considered to be insensitive towards small changes of the input parameters. Moreover, they do not suffer from the curse of dimensionality. The worst performance corresponds to the 50%-50% splits of the RBF network. In this case, F_1 measure equals 0.885. This can be due to that there are several parameters to be tuned, such as minimum standard deviation of the Gaussians, which can easily lead to misclassification.

The 70%-30% split demonstrates improved performance compared to the 50%-50% split for all the neural networks but the 3-layered PSO-trained MLP feed-forward network. This may be ascribed to the fact that the PSO-trained MLP network for the 70%-30% split tends to overfit the data. Concerting the accuracy range between the 70%-30% split and the 50%-50% split, it is easy to deduct that the neural network with the greater range is back-propagation MLPs. In back-propagation MLPs the accuracy range is 4.80. VPs are the least affected. In this case the accuracy range is 2.09.

6. Conclusions

A novel framework for audio dialog detection is described. The cross-correlation of a pair of indicator functions and the corresponding cross-power spectral density are utilized as inputs to a variety of neural networks, namely: perceptrons (MLPs, 3-layered PSO-trained MLPs, VPs), radial-basis function networks and support vector machines. Audio scenes are extracted from six movies, containing dialogs and non-dialogs. SVMs with second order polynomial kernel achieved the highest average dialog detection accuracy, equal to 94.740%. All results are superior to state-of-the-art dialog detection techniques [12]. However, since there is no common database for performance evaluation, direct comparison is not feasible.

Acknowledgement

This work has been supported by the research project 03ED849. E. Benetos was supported by the "Alexander S. Onassis" Public Benefit Foundation through scholarship.

References

[1] A. A. Alatan, A. N. Akansu, and W. Wolf, "Comparative analysis of hidden Markov models for multi-modal dialogue scene indexing," in Proc. *2000 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2000, vol. 4, pp. 2401-2404.

[2] A. A. Alatan and A. N. Akansu, "Multi-modal dialog scene detection using hidden-markov models for content-based multimedia indexing," *J. Multimedia Tools and Applications*, 2001, vol. 14, pp. 137-151.

[3] D. Arijon, "*Grammar of the Film Language*," Silman-James Press, 1991.

[4] B. Birge, "PSOt - A particle swarm optimization toolbox for Matlab," in Proc. *2003 IEEE Swarm Intelligence Symp.*, 2003, pp. 182-186.

[5] L. Chen and M. T. Özsu, "Rule-based extraction from video", in Proc. *2002 IEEE Int. Conf. Image Processing*, 2002, vol. 2, pp. 737-740.

[6] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in Proc. *6th Int. Symp. Micro Machine and Human Science*, 1995, pp. 39-43.

[7] Y. Freund and R. E. Schapire, "Large margin classification using the perceptron algorithm," *Machine Learning*, 1999, vol. 37, no. 3, pp. 277-296.

[8] D. W. Hosmer and S. Lemeshow, "*Applied Logistic Regression*," N.Y.: Wiley, 2000.

[9] G. Iyengal, H. J. Nock, and C. Neti, "Audio-visual synchrony for detection of monologues in video archives," in Proc. *2003 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2003, vol. 1, pp. 329-332.

[10] M. Kotti, E. Benetos, and C. Kotropoulos, "Automatic speaker change detection with the bayesian information criterion using MPEG-7 features and a fusion scheme," in Proc. *2006 IEEE Int. Symp. Circuits and Systems*, 2006, pp. 1856-1859.

[11] M. Kotti, C. Kotropoulos, B. Ziólko, I. Pitas, and V. Moschou, "A framework for dialogue detection in movies," *Lecture Notes in Computer Science*, 2006, vol. 4105, pp. 371-378.

[12] P. Král, C. Cerisara, and J. Kleckova, "Combination of classifiers for automatic recognition of dialogue acts," in Proc. *9th European Conf. Speech Communication and Technology*, 2005, pp. 825-828.

[13] B. Lehane, N. O'Connor, and N. Murphy, "Dialogue scene detection in movies using low and mid-level visual features," in Proc. *Int. Conf. Image and Video Retrieval*, 2005, pp. 286-296.

[14] L. Lu and H. Zhang, "Speaker change detection and tracking in real-time news broadcast analysis," in Proc. *2004 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2004, vol. 1, pp. 741-744.

[15] S. Nissen, "Implementation of a fast artificial neural network library (FANN)," *Technical Report*, Department of Computer Science, University of Copenhagen, Denmark, October 2003.

[16] J. Platt, "Fast training of support vector machines using sequential minimal optimization," in B. Schoelkopf, C. Burges, and A. Smola, eds., *Advances in Kernel Methods - Support Learning*, MIT Press, 1999.

[17] P. Stoica, R. L. Moses, "*Introduction to Spectral Analysis*," Upper Saddle River, NJ: Prentice Hall, 1997.

[18] I. C. Trelea, "The particle swarm optimization algorithm: convergence analysis and parameter selection," *Information Processing Letters*, 2003, vol. 85, pp. 317-325.

[19] A. Vassiliou, A. Salway, and D.Pitt, "Formalising stories: sequences of events and state changes," in Proc. *2004 IEEE Int. Conf. Multimedia and Expo*, 2004, vol. 1, pp. 587-590.

[20] Y. Zhai, Z. Rasheed, and M. Shah, "Semantic Classification of Movie Scenes Using Finite State Machines," *IEE Proc. - Vision, Image, and Signal Processing*, 2005, vol. 152, no. 6, pp. 896-901.

[21] MPEG-7 overview (version 9), *ISO/IEC JTC1/SC29/WG11 N5525*, March 2003.