Multi-Faceted Contextual Model for Person Identification in News Video

Ming Zhao, Shi-Yong Neo, Hai-Kiat Goh, Tat-Seng Chua
School of Computing, National University of Singapore
{zhaom, neoshiyo, gohhaiki, chuats}@comp.nus.edu.sg

Abstract

Person identification is very important in the domain of multimedia news as it is often the focus of events in news stories and interest of searchers. However, this detection is impeded by the imprecise audio/visual analysis tools. In this paper, we describe a multimodal and multi-faceted approach to Person-X detection in news video. We make use of multimodal features extracted from text, visual and audio inherent in news video. We also incorporate multiple external sources of news from web and parallel news archives to extract location and temporal profile of the persons. We call this second source of information the multi-faceted context. The multimodal, multi-faceted information is then fused using a RankBoosting approach. Experiments on TRECVID 2003 and 2004 search queries demonstrate that our approach is effective.

1. Introduction

Digitalized videos are now widely available and can be accessed using any computer or home entertainment system. With high-capacity storage devices and ever-increasing amount of information on the web, we are currently facing a situation where it is almost impossible to find what we want by browsing through the vast amounts of video information manually. This has resulted in a need for an effective video retrieval system that is capable of returning multimedia results that match user's search interests. Our work targets at news video, and in particular, the detection and recognition of people in the video footages known as Person-X detection [3,13,15]. As news are concerned mostly with news-worthy subjects involving various people, the ability to detect and recognize people is essential to better indexing and access of news video contents. In particular, it supports high-level tasks such as finding the events involving a named person, or probing the name and identity of an unknown face.

Current work on Person-X detection focuses mainly on fusing various multimodal features. The features used include automatic speech recognition transcript (ASR), video optical character recognition (OCR), shot boundary, audio class, as well as derived features such as face, anchor-person, shot genre, speaker change segment, name entity at the shot level. Although such systems are effective, they often miss faces where names are not referenced explicitly. One such example is "Madeleine Albright" appearing in a live footage of a summit with "Bill Clinton" but her name is not mentioned. To capture such knowledge, we need to bring in temporal and location context. Thus in this research, we extend the multimodal Person-X detection model to incorporate the multi-faceted model of context in both time and space dimensions. Such time-dependent spatial knowledge can be derived by analyzing the "person to person" and "person to location" correlations from external information sources such as news reporting of the corresponding person. The multi-faceted correlation knowledge is then fused with the various multimodal feature of news video using a variety of fusion approach, including the RankBoost technique [5]. The resulting model has been found to outperform conventional multimodal based approach to Person-X detection when tested on TRECVID 2003 and 2004 datasets. The following sections describe the details of our approach.

2. Related Work

Given a news video, it is ideal if we can actually name all the persons that appear in the video automatically (Person Naming). However, this is clearly not achievable with the current technology. Even manual annotators will not be able to name every single person without substantial prior knowledge. Therefore, a more practical approach is to detect only persons that are required by the users (Person Finding).

In order to search for a certain person appearance in news video, one of the simplest approaches is to look for the person's name in the ASR. The Name-It [15] uses this approach to associate names and faces based on their co-occurrences in the news video. A face image is labeled with the name that has the largest temporal overlap with a group of images containing faces similar to the given one. With a given name, the face images are found in a similar way. This method can thus be used for both person naming and person finding. Houghton et al. [13] built a database of named faces to allow users to query a person's name by giving an image containing the target face. The named faces are found when the names are overlaid on video frames and recognized by video OCR. Albiol et al. [1] used speaker recognition and face recognition technologies to find a specific person in news sequences. In contrast, Chen et al. [3] employed text, timing, anchor person and face information to find a specific person. They used the text information to search the shots mentioning the person's name, and used time relationship between the mentioning of the name and the appearance of the person to propagate the text search results to nearby shots. Anchor person detection and face recognition are also used to further improve the result. The final prediction is based on the linear combination of the text information, anchor person detection and face recognition. Yang et al. [21] attempted to name every individual person appearing in the broadcast news videos with names detected from the video transcript. It used two kinds of information: features and constraints. A learning method is used to predict for each person the most likely name based on the features and constraints.

In general, existing methods for person-X detection focus on optimizing the use of specific modal features but do not leverage on the use of external resources. It therefore limits the performance of systems which are sensitive to the accuracy of various feature extraction techniques.

3. Multimodal Features

Based on previous review, we intend to perform Person-X detection by using multimodal features as well as contextual knowledge derived from external information sources. In our work, we plan to make use of 3 categories of multimodal features. They are: text, visual and audio.

3.1. Text

The textual features used include: (a) ASR output [7] and (b) video OCR output [3]. Generally, the appearance of Person-X is correlated to the appearance of the name in ASR or Video OCR. It is, however, possible that the appearance may not coincide with the shots which the name is mentioned. In addition, there are cases where the name may not be mentioned in the ASR or is missing due to ASR errors. On the other hand, although video OCR provides very good indication of the appearance of Person-X, most occurrences of faces do not have corresponding video text. Thus the two main problems in utilizing ASR and OCR for name matching are: (1) the recognition error in OCR, and (2) temporal mis-alignment in ASR in which the face usually does not show up in the shot where the name is mentioned.

To overcome the first problem, we employ minimum edit distance (MED) to correct the insertion, deletion and mutation errors in OCR. Let S_{Name} be the name of Person-X and S_{OCR} be an OCR word, the minimum edit distance between them is:

$$Dist_{MED} = MED(S_{Name}, S_{OCR}) \qquad (1)$$

where *MED* is the minimum edit distance function. If $Dist_{MED}$ for any OCR word of a particular shot is smaller than a threshold α (determined by the length of word), the OCR is identified to contain S_{Name}.

To overcome the second problem, we use the distribution statistics collected from the training data that models the probability between the shot where the name appears and those that the face actually occurs. We use this distribution probability to propagate the similarity scores from the shots containing the person's name to the neighboring shots in a window. The propagation is carried out as:

$$Sim_p(X,S) = \sum_{|S-S_i|<w} p(S,S_1) Sim_p(X,S_i) \qquad (2)$$

where w is the size of the window measured either by time or by shot offset, S_i is the neighbor shots and $p(S,S_1) \in [0,1]$ is a propagating function, which determines how the similarity is propagated to neighboring shots. For a specific person, we estimate a Gaussian distribution using the temporal distances from the appearances of names to the corresponding visual appearances. Besides using name of the Person in ASR, we also make use of other keywords that co-occurs with the name as extra contextual terms as discussed in *Section 4*.

3.2. Visual

Visual information provides valuable clues for finding a person in news video. Unlike text information which roughly estimates where the person is, visual information is able to pin-point the exact position and time of the person's appearance. We consider two types of visual features: face recognition and anchor person detection. Face recognition identifies a known person's face from a database of images of known persons. Anchor person detection helps to provide partial semantics for news segmentation as well as a guide for filtering irrelevant shots. Our face recognition and anchor person detection systems are based on the techniques describe in [4].

3.3. Audio

Audio information such as speech and speaker recognition is another important feature for news video. It provides time and possible event information. Besides utilizing audio for ASR, we also make use of audio to identify specific audio classes and speakers. We have identified several audio classes that are important to Person-X detection: silence, music, female speech, male speech, and noise. We employ Mel-frequency Cepstrum Coefficient (MFCC) together with zero crossing rate, centroid and roll off point energy as features [9,10]. These features are used for audio classification using a combination of K-means classifier and multidimensional HMM training model. To perform speaker identification, we first divide the continuous speech segments using speaker change information [7]. Subsequently, each segment is matched with the training samples of each speaker to obtain a probability of the speaker as in [4].

4. Additional Context from External Sources

With the availability of online news articles, it is possible to obtain related news articles to complement the short and summarized versions in news video. We have chosen three external sources of information: (a) AQUAINT text corpus, (b) news websites, and (c) web search engines. The AQUAINT corpus contains news articles from the same period as the test video, whereas the other two sources contain latest information which is not from the same period. Given the name of a person-X to be identified, our aim is to induce additional contextual knowledge from these external information sources. The knowledge could come from keywords, phrases and named entities (NE) and their correlations in time and space. Such knowledge is termed multi-faceted contextual knowledge as it covers correlations of names, locations, dates, events etc. For ease of discussion, we use the term Contextual Terms (CTs) to collectively denote these additional entities.

4.1. Contextual Terms

Let the set of Person-X to be detected be $\{E_i\}^{i=1:n}$. Given the set of articles from training transcripts and external information sources, we first perform stop-words filtering and named entity extraction [20]. We then compute the mutual information gain [6,11] between potential CTs and Person-X as follows:

$$\begin{aligned} G(CT_x) = \omega_x \times [-\sum_{i=1}^{n} \Pr(E_i) \log \Pr(E_i) \\ + \Pr(CT_x) \sum_{i=1}^{n} \Pr(E_i \mid CT_x) \log \Pr(E_i \mid CT_x) \\ + \Pr(\overline{CT_x}) \sum_{i=1}^{n} \Pr(E_i \mid \overline{CT_x}) \log \Pr(E_i \mid \overline{CT_x})] \end{aligned} \quad (3)$$

We use a threshold β to select the initial set of CT_x.

As online news websites and search engines mostly index recent articles that may not correspond to our video data time range, we assign higher weights ω_x to CTs obtained from the AQUAINT corpus which corresponds to our video dataset. Individual CT, however, may not be a good indicator of a given person-X or NE. For example, an NE "White House" may be used to infer both "George W. Bush" and "Richard B. Cheney". However, "White House" and "president" can be used together as a strong indicator of "George W. Bush". Thus related CTs are further grouped according to their confidence in inferring person-X. The confidence for a group of CTs to infer person-X or NE is given as follows:

$$C(\text{Person-X} \mid \{CT_x\}_{x=1}^{n}) = {}^{1}\!/_{n} \sum_{x=1}^{n} \Pr(\text{Person-X} \mid CT_x) \quad (4)$$

The final lists of CTs are then time-stamped according to the estimated time and date of the documents they appeare in. These CTs are utilized in the retrieval of videos at a later stage and also in the following correction task.

4.2. Correction of Name Entities in ASR

ASR is erroneous partly due to the complexity of human speech. The error is particularly serious for Name Entities involving non-Latin-based person names. To improve the performance, we make use of the CTs extracted online to perform correction on the

ASR. The correction algorithm is based on the phonetic similarity as well as the story context [20]. Two similarity measures are used for matching the CTs with ASR texts: a) Phonetic String Boundary (PSB), which is assigned a higher score if the starting and ending phonetic sounds [16] match; b) Longest Common Phonetic Subsequence (LCPB). We only use CTs having time-stamps within a range ±*T* of the video data dates to correct the relevant ASR text or NEs.

4.3. Locality as Time Dependent Context

Most Person-X detection systems rely on the story context from the ASR for initial retrieval and subsequently make use of audio/visual features to pinpoint possible shots which the person may appear in. In this research, we introduce another important facet, locality as a form of *time dependent context*. For example, a user looking for "Sam Donaldson" in 1998 can generally use the location "White House" to associate him, since Sam is a correspondent for White House during that time period. Often, humans tend to associate people to various locations at different time when we are doing a search. Therefore, we deem the locations where person-X frequently appears as an area of influence for that person. The area of influence provides indicative locations for person-X appearance. This is reasonable since it is quite unlikely that Sam will appear in countries like India in 1998 news.

There is an additional retrieval advantage if we know where some news is happening and where they are being reported. For very famous people like "George Bush", we can track their "has-been" locations by analyzing news articles which are related to him. For instance, if George Bush is visiting Britain today, there will likely be news articles describing his visit. Such location information can be gathered over a period of time and use as a journey log of that person. This information can be derived from various external resources.

We propose the use of parallel corpus to obtain the context and area of influence (AOF) of any given person. The AOF is formally defined as follows: *The effective context or location in the real world that a given person is most likely to appear within a specific time range.* We propose to build the AOF by using the set of related news articles instead of directly from ASR text from video. The reasons that we rely on external news articles are: (a) they are relatively free of errors; and (b) they tend to have more descriptive or lengthy statements as compared to ASR text. Given an article related to Person-X at time *t*, we obtain the lists of location NEs (L_j) and person NEs (P_k) by using a NE extractor [20]. With a collection of related news article over a period of time, we can derive the probability of "Person-X" occurring with other person, or in certain location with respect to time. For example, if there are text articles at $Date_t$ describing "Bill Clinton" and "Hillary Clinton" giving a speech in Israel's summit meeting. Then there is a high probability that (Location "Israel" +Time "*t*") → "Bill Clinton", "Hillary Clinton". Therefore we can calculate the co-occurrence probability of "Bill Clinton" at a given time *t* with respect to other locations and people as follows:

$$P(C_X \mid t, L_j, P_k) = P(C_X \mid \Theta^X_{t,L_j,P_k}) \qquad (5)$$

The model Θ^X is built from related text news articles of Person-X. The co-occurrence probability C_X can suggest if a particular news segment is relevant to Person-X even if his name is not mentioned in the ASR. C_X, however, needs to be assigned correctly to a suitable segment in the video in order to make sense. As the short ASR text at the shot level may not contain the desired information, a story-level unit is necessary for contextual analysis. Here, we employ a multimodal HMM based technique to extract the story boundaries [2]. Given a story level segment at time t, we can determine C_X by using the name entities of type locations (L_j) and person (P_k) extracted from the ASR as:

$$P(C_X \mid Story_Y) = \arg\max_{tjk}(P(C_X \mid \Theta^X_{t,L_j,P_k})) \qquad (6)$$

5. Methods of Fusion

The fusion process is the most important part of a Person-X detection system. It is clear that none of the features alone is able to provide sufficient semantics in detecting Person-X accurately. Therefore, it is necessary to develop an appropriate way to fuse all information from various modalities. The modalities (or features) we use are

ASR, NEs within the segments, video OCR, audio classes, faces, face recognition, anchor-person, shot-genre, speaker changes, speaker identification as well as the locality. In this research, we investigate 2 effective techniques for fusion: (1) scoring fusion and (2) rank-list fusion.

5.1. Scoring Fusion

This fusion process combines the separate scores from different features into one fusion score. The separated scores can be confidences, probabilities, similarities or distances depending on the methods used for each feature. Training and test samples are then ranked according to the final fusion score. We employ two scoring fusion methods in this paper: linear weighted sum and SVM fusion.

a) Linear Weighted Sum. One of the properties of sum scoring is that it does not magnify noise as severely as product scoring. This ensures that the linear weighted sum method is more robust to noise. Given scores *(S_1, S_2, ..., S_n)* from different features for a sample x, the linear weighted sum fusion is:

$$F_{LWS}(x) = \sum_{i=1}^{i=n} w_i * s_i \tag{7}$$

where w_i is the weights trained based on the training data. In our work, we use gradient descent [12] to find the optimal weights for maximizing the average precision.

b) Non-linear Weighted Sum (pSVM). As linear weighted sum is a form of linear model, it is not capable of modeling the inter-dependencies between features. A theoretical upper bound for the average precision of the linear combination is presented in Yan et al. [19]. They showed that the linear combination method has serious theoretical limitations. Although linear combination might be sufficient for a small number of features, they suggested that non-linear combinations should be used to make use of inherent relationships between features for a large feature set. As such, another fusion method that we utilize is probabilistic SVM (pSVM). In our implementation, we make use of radial basis function (RBF) kernel and logistic regression for computing the probabilities [14].

5.2. Rank List Fusion

Rank list fusion [5] mainly consists of two steps: (a) ranking results of different features according to their scores; (b) fusing the multiple rank lists into one rank list. Rank list fusion differs from scoring fusion in that it concentrates on the ranking of items based on each feature rather than the actual score fusion. As different scores from different features may have different meanings, it is inappropriate to merge them without considering their common basis. Rank list fusion avoids this problem by only utilizing the relative scores from rank lists of each feature type.

We employ Rankboost [5], which is a powerful algorithm for combining rank lists or preferences, for our rank list fusion. It has been successfully employed in information retrieval, natural language processing and shape localization. Let ***X*** be a set called the *instance space* and $x \in X$ be the *instances* we are interested in, the RankBoost algorithm is shown in Figure 1.

```
Algorithm RankBoost
1. Given initial distribution $D_0$ over $X \times X$
2. Initialize $D_1 = D_0$
3. For $t$ = 1, …, T:
    • Train weak learner using distribution $D_t$
    • Get weak ranking $h_t : X \to R$
    • Choose $\alpha_t \in R$
    • Update
        $D_{t+1}(x_0, x_1) = \frac{D_t(x_0, x_1)\exp(\alpha_t(h_t(x_0) - h_t(x_1)))}{Z_t}$
        where $Z_t$ is a normalization factor
        (chosen so that $D_{t+1}$ is a distribution)
4. Output the final ranking
        $H(x) = \sum^{T} \alpha_t h_t(x)$
```

Figure 1. **Algorithm for RankBoost**

In order to combine separate rank lists into a single final rank list based on feedback information, RankBoost builds a strong ranking function H_T from T weak ranking functions h_t. The weak ranking function h_t is derived from a rank list f_n by comparing the score of f_n on a given instance to a threshold μ. For instances unranked by f_n, the weak ranking function assigns a default score ε. In-depth descriptions of the algorithm can be found in [5]. With respect to Person-X detection, the rank lists f_n corresponds to the rank lists obtained from detectors which uses different features. The weak ranking functions h_t are subsequently acquired using these rank lists f_n. At the final stage, the ranking function H_T combines all weak ranking functions h_t to form a final rank list. The algorithm using a two-level fusion model is shown in Figure 2.

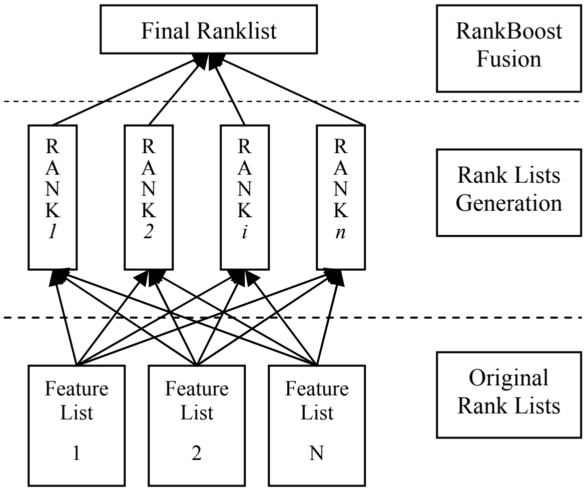

Figure 2.**Two-level RankBoost fusion model**

We propose this two-level fusion model to allow for a better modeling of relations between features. For example, if OCR and face recognition detectors output a high rank for a specific image, it is very likely that this image contain Person-X. However, it is possible that this image is being ranked lowly by all the other detectors. If we use a one level fusion, this may offset the good results provided by some detectors combinations. Hence it may cause the image to have an overall low rank. To resolve this, we create multiple 2nd level rank lists by combining results from different feature detectors. These intermediate rank lists can provide a stronger indication of which image samples may be positive. We further devise 2 methods to combine these intermediate rank lists: Intersection and Union.

Intersection generation only combines the samples that appear positive in two or more detectors. This method is used to generate rank lists with high precision. A sample which is classified as highly positive by many detectors is more likely to be a true positive. Using this method, rank lists with high precision can be selected by Rankboost to assign their respective samples a higher rank in the final rank list. Intersection generation can be implemented by both linear weighted fusion and SVM fusion.

Union generation combines all samples that appear positive in at least one detector. This method is used to generate results with high recall. There are situations where some features by themselves are very accurate. For example, the speaker recognition is very accurate for detecting the presence of a speaker. However, speaker presence may not be detected by using face, OCR or ASR at times. In such cases, intersection generation will miss these positive samples. On the other hand, union generation is able to collect these missed positive samples and form rank lists with high recall. Similar to intersection generation, it can also be implemented by both linear weighted fusion and SVM fusion.

6. Experiments

To test the effectiveness of our approach, we have chosen 10 people to be detected. This selection is based on the queries of TRECVID 2003 and 2004 search task [17]. In the search task, the participants are required to submit a ranked list of shots, given a multimedia query. The query consists of a short text description and may be accompanied by short video clips and/or images. The 10 chosen people are listed in Table 1. We divide the 60 hours of videos from October 1998 to December 1998 from CNN and ABC broadcast equally for training and testing. We follow the performance measure used in TRECVID, which is mean average precision (MAP). This performance measure is widely used for system evaluation in information retrieval over large corpuses where the recall rate is difficult to determine. The system will return at most 2000 shots which are deemed to contain the desired Person.

Table 1. **List of 10 people for detection**

Yasser Arafat	Bill Clinton
Pope John Paul II	Sam Donaldson
Mark Souder	Saddam Hussein
Morgan Freeman	Boris Yeltsin
Henry Hyde	Benjamin Netanyahu

6.1 Text-based Vs Multimodal

The first series of tests is designed to evaluate the following two premises. First, we want to know the performance of using text-only features. This will serve as the performance baseline. Second, we want to determine the amount of improvement by adding the rest of the multimodal features as described in *Section 3*. We carry out 3 runs as follows:

P0) Baseline (basic text retrieval from ASR)
P1) P0 with Contextual Terms (derived from relevant news articles)
P2) P1 with other multimodal features combined using linear-weighted sum

Table 2. **Results in terms of MAP**

Runs	P0	P1	P2
MAP	0.16	0.21	0.33

From Table 2, it is clear that the use of contextual terms extracted using Eqn (4), in addition to Person's name, is helpful to Person-X detection. We observe a significant improvement in performance from P0 to P1. Further, the use multimodal feature is essential as the MAP increases by about 50% from P1 to P2. This can be attributed to the use of Video OCR and face detector. The major jump in precision is attributed to a better ranking of shots as textual features alone are not sufficient to determine whether a face is present in a shot. The filtering of anchor-person shots and other shot-genres like commercial, text scenes, weather news also increases the overall performance. In addition, video OCR also improves the precision further by providing precise timing information of the appearance of the Person's name on the screen.

6.2 Effects of Multi-faceted Contextual Information and Fusion Schemes

Next, we design experiments to demonstrate the effectiveness of contextual information and various fusion methodologies. We create 4 runs with the following objectives: a) to determine the effects of multi-faceted context; and b) to find the best fusion method so that the system can achieve the best MAP score. The 4 runs are as follows:

P3) multimodal features with SVM
P4) multimodal features with RankBoost
P5) P3 with Time Dependent Context extracted using Eqn (6)
P6) P4 with the same Time Dependent Context as P5

Table 3. **Results in terms of MAP**

Run	P3	P4	P5	P6
MAP	0.34	0.38	0.38	0.41

From the results, we observe that RankBoost fusion (P4) is better than SVM (P3) as well as linear-weighted sum fusion (P2). The increase in MAP is due to better ranking of shots. In RankBoost fusion, each shot is not only denoted by its feature score, but also by its ranking position. If given a feature where the scores of shots are almost equal (i.e. the difference between top ranked shot and the top 100th shot is very small), score fusion will likely to return almost equal score for that feature. However, RankBoost will be able to give a better judgment. We also see that the use of non-linear scoring fusion (P3) only increases the performance slightly as compared to linear weighted sum (P2). This is because that the number of features is small. As discussed by Yan et al [19], the linear weighted sum is good for a small number of features.

In addition, we observe that runs P5 and P6 that make use of multi-faceted contextual information performs much better than those without. This shows that the time dependent context is effective. We also find that this multi-faceted information works well for very famous people like Bill Clinton, Boris Yeltsin, Pope John Paul II and Benjamin Netanyahu as there are plenty of related news articles. For example, the system is able to find shots of Bill Clinton even when his name is not mentioned or he is facing sideway where face detection is not effective. We are also able to find shots where Pope John is standing far away where his face is not detected.

7. Conclusions and Future Works

This paper discussed the techniques employed in our Person-X detection framework. We introduced the use of multi-modal and multi-faceted context involving time and location. We then fused this multimodal, multi-faceted information using various methodologies including RankBoost. From the experiments, we see that the multi-faceted approach complements the retrieval by providing valuable information that is not available in audio–visual content and surrounding ASR text.

For future work, we will explore better model to incorporate contextual information to support Person-X detection. In addition, we will explore other fusion methods using model-based approach like Hidden Markov Random Fields (HMRF).

8. References

[1] Albiol, A.; Torres, L.; Delp, E.J. "The indexing of persons in news sequences using audio-visual data", *IEEE International Conference on Acoustics, Speech, and Signal Processing,* 2003. Volume 3, 6-10 April 2003 Page(s):III137-40.

[2] L. Chaisorn, T.-S Chua and C.-H. Lee. "The segmentation of news video into story units." *IEEE Int'l Conf. on Multimedia and Expo*, 2002.

[3] M. Y. Chen and A. Hauptmann. "Searching for a specific person in broadcast news video." *Proc. of the Int'l Conf on Acoustic, Speech and Signal Processing*, Vol. 3, 1036-1039. May 2004.

[4] T. S. Chua, S. Y. Neo, K. Y. Li, G. Wang, R. Shi, M. Zhao and H. X. Xu, "TRECVID 2004 Search and Feature Extraction Task by NUS PRIS". In *TRECVID* 2004, NIST, Gaithersburg, Maryland, USA, 15-16 NOV 2004.

[5] Y. Freund and R. E. Schapire, "A Decision-theoretic generalization of online-learning and an application to boosting". *Journal of Computer and System Sciences*, Vol. 55, no. 1, 119-139, August 1997.

[6] Ciravegna, F. 2001. "An Adaptive Algorithm for Information Extraction from Web-related Texts". *In Proc. of IJCAI-2001 Workshop*.

[7] J.L. Gauvain, L. Lamel, and G. Adda. "The LIMSI Broadcast News Transcription System". *Speech Communication*, 37(1-2): 89-108, 2002.

[8] A. Hadid and M. Pietikainen, "Selecting models from videos for appearance-based face recognition," *in Proceedings of International Conference Pattern Recognition*, August 2004, vol. 1, pp. 304– 308.

[9] A Hauptmann, R. Jin and T. D. Ng. "Video Retrieval using Speech and Image Information". *Proc. of Electronic Imaging Conference (EI'03)*, Storage and Retrieval for Multimedia Databases, Santa Clara, CA, Jan 2003.

[10] H. Jiang, T. Lin and H.J. Zhang. "Video segmentation with the Support of Audio Segmentation and classification". *ICME'2000-IEEE Int'l Conf on Multimedia and Expo*, NY, USA, Jul 2000.

[11] C. Kenneth and P. Hanks. "Word Association Norms, Mutual Information, and Lexicography". *Proc. of the 27th Annual Meeting of the ACL*, 1989.

[12] J. Kittler, M. Hatef, and R. P. W. Duin. "Combining classifiers". *Intl. Pattern Recognition*, pages 897–901, 1996.

[13] Houghton, R. Named Faces: "Putting Names to Faces". *In IEEE Intelligent Systems Magazine*, 14(5): 45-50, 1999

[14] J. Platt. "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods". *Advances in Large Margin Classifiers*, MIT Press, pages 61–74, 2000

[15] S. Satoh and T. Kanade, "Name-it: Association of faces and names in video," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997, pp. 368–373.

[16] K. Seymore, S. Chen, S. Doh, M. Eskenazi, E. Gouvea, B. Raj, M. Ravishhankar, R. Rosenfeld, M. Siegler, R. Stern, and E. Thayer (1998). "The 1997 CMU Sphins-3 English Broadcast News Transcription System",. *Proceedings of the 1998 DARPA Speech recognition Workshop*

[17] TRECVID, *TREC Video Retrieval Evaluation*, http://www-nlpir.nist.gov/projects/trecvid

[18] R. Yan, J. Yang, and A. G. Hauptmann. "Learning Query-Class Dependent Weights for Automatic Video Retrieval". *Proc. of ACM MM*, New York, Oct 2004.

[19] R. Yan and A. G. Hauptmann. "The combination limit in multimedia retrieval". *ACM Multimedia*, 2003.

[20] H. Yang, L.Chaisorn, Y. Zhao, S.Y. Neo, T.S. Chua, "VideoQA: question answering on news video", *Proc. of ACM MM*, Berkeley, 632-641, Nov 2003.

[21] J. Yang, M. Chen, A. G. Hauptmann: "Finding Person X: Correlating Names with Visual Appearances". *CIVR 2004*: 270-278

[22] W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, December 2003