Multi-modal Analysis for Person Type Classification in News Video

Jun Yang, Alexander G. Hauptmann

School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave, PA 15213, USA

{juny, alex}@cs.cmu.edu, Tel: 412-268-1448, Fax: 412-268-5576

ABSTRACT

Classifying the identities of people appearing in broadcast news video into anchor, reporter, or news subject is an important topic in high-level video analysis. Given the visual resemblance of different types of people, this work explores multi-modal features derived from a variety of evidences, such as the speech identity, transcript clues, temporal video structure, named entities, and uses a statistical learning approach to combine all the features for person type classification. Experiments conducted on ABC World News Tonight video have demonstrated the effectiveness of the approach, and the contributions of different categories of features have been compared.

Keywords: Multi-modality analysis, person type classification, broadcast news video

1. INTRODUCTION

The activities of people play a primary role in conveying the semantics of videos of various genres. Therefore, detecting people's appearances and recognizing their identities and roles in the video is of great research value for better video indexing and access. This is of particular interest to broadcasting news video since it involves a large number of important people. Therefore, there have been extensive works on people-related tasks in news video analysis, such as detecting faces/people from the video frames [5], finding the appearances of a named person [12], correlating people names in closed-captions with faces detected from video frames [8,13], creating a face database from news video to allow users to query the name of an unknown face image [4]. This paper focuses on a relatively understudied sub-problem, namely person type classification in broadcast news videos, which intends to classify each person appeared in the video into three types as:

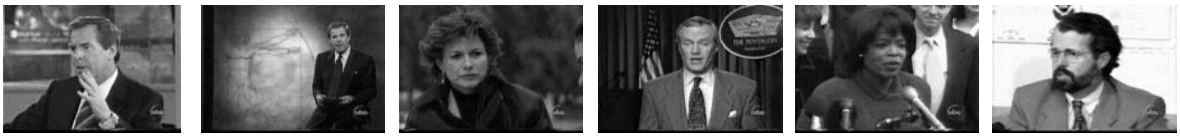
(1) Anchor, the person who narrates or coordinates the news broadcast;

(2) Reporter, the person who reports the details of a news event; and

(3) News-subject, the person involved in a news event.

A news video can be partitioned into a series of news stories with each story consisting of several camera shots [14]. In a typical news story, people of the three aforementioned types may appear, usually including one anchor, one or two reporters, and a varying number of news subjects. Despite the extensive work on news video analysis, classifying a person's type as anchor, reporter, and news-subject remains a missing piece, which is important to many other tasks. For example, once a person's type is known, we can eliminate anchors and reporters as false alarms when trying to find a named news-subject, or predict a person's name with higher accuracy. For example, when searching for a named news-subject, the results are easily mixed up with shots of anchors and reports, which can be eliminated if a person’s type can be predicted accurately. Moreover, the knowledge on a person’s type will help the prediction of his/her name given that the types of all the people names are also known (which is relatively easy).

Figure 1 shows the examples of anchors, reporters, and news-subjects in a news video, which are quite difficult to distinguish visually, especially between reporters and news-subjects. As we have observed, there is unlikely to exist a "magic" features that can tell them apart accurately; nevertheless, there exist many weak clues from multiple modalities of video that imply the type of a person. For example, on the visual aspect, anchors and reporters usually have frontal faces, while news-subjects may have side faces; on the audio aspect, anchors and reporters are faster speakers than news-subjects; on the text aspect, there are certain “clue phrases” that the anchors and reporters introduce themselves and greet the audience. Therefore, selecting discriminative features from multi-modal analysis and combing them effectively is essential to the success of person type classification.



Anchor Anchor Reporter Reporter News subject News subject

Figure 1: Examples of anchors, reporters, and news subjects

For simplicity, our work focuses on persons who are giving monologue-style speech individually (alone) in the video. This however does not cost much generality of this work, given the observation that (1) anchors and reporters appear individually in most cases and thus multiple people appearing in the same video frame are probably all news-subjects, and (2) people rarely appear in a news story without speaking at some point. With this simplification, our work boils down to classifying *monologue shots* (i.e., video shots where someone is delivering a monologue speech) into anchor shot, reporter shot, or news-subject shot. In this paper, we assume that all the monologue shots have been identified manually or using automatic approaches [10].

2. MULTI-MODAL VIDEO ANALYSIS

2.1. Feature selection methodology

Features used for a high-level video analysis task are very *ad-hoc*. The features used for, say, sports news detection, can be very different from those used for commercial detection. Nevertheless, we argue that the process of discovering and selecting features is somewhat similar among different tasks. As shown in Figure 2, the feature selection consists of a forward process and a backward one. Using person type classification as an example, in the forward process, we manually label some shots as anchor, reporter, and news-subject, inspect them in order to find out features useful for discriminating shots of different types, and then train a classifier based on these features. In the backward process, we apply the trained classifier on the labeled data, analyze the classification errors so that noisy features causing the errors are removed and additional features useful for correcting the errors are included. The classifier is then re-trained based on the updated features. The backward process is like a "feedback" process which is repeated until the update of the feature set no longer reduces misclassifications significantly. Though somewhat labor-intensive, selecting effective features is critical to the success of any video analysis task.

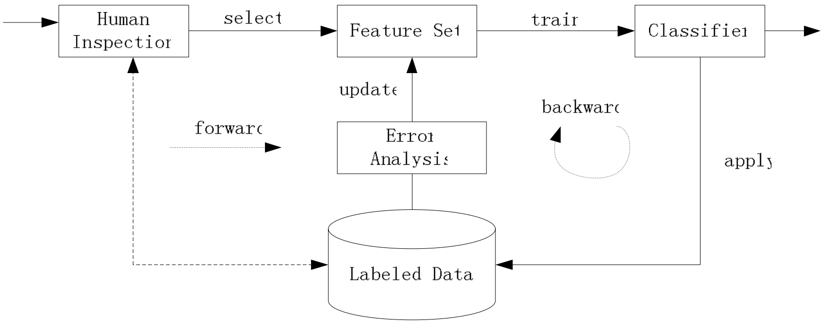


Figure 2: Feature selection methodology

2.2. Multi-modal features

2.2.1. Transcript clues

In our work, video transcript is the closed-caption which has been temporally aligned to the video. In the transcript of each type of broadcasting news, there are certain “clue phrases” that allow us to identify reliably some of the anchor and

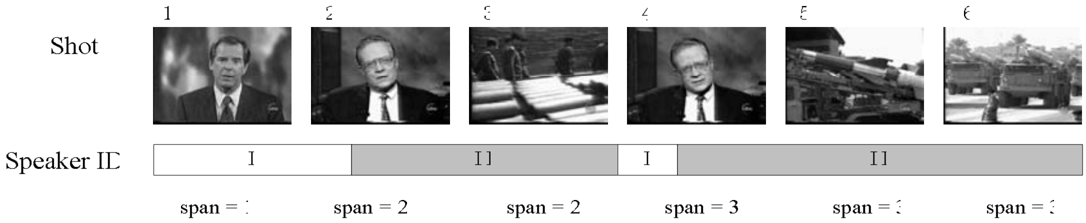


Figure 3: Speaker identification analysis

reporter shots. For example, an anchor of ABC World News Tonight normally uses one of a few fixed ways to introduce a reporter, such as "*ABC's Linda Douglass has the story*", and to conclude one day's broadcasting, such as "*I'm Peter Jennings*. *Have a good night*". The occurrences of such clue phrases indicate anchor shots. Similarly, a clue phrase indicating a reporter's shot is the self-introduction of reporters, such as "*Barry Serafin, ABC news, in Washington*". Since there are only a small number of clues, we handcraft them as several “templates”, which are used to find all their occurrences in the transcript automatically. Detecting some clue phrases require the identification of people names, which will be discussed in the next subsection.

2.2.2. Name entities

There exist many people names in the video transcript. Although the associations between the names and the people (shots) are unknown, the people names detected from the transcript still provide valuable clues for identifying a person's type. For example, if no reporters' names are spotted in the transcript of a news story, there is probably no reporter appearing in this story. We apply the well-known BBN's named entity detection algorithm [1] to extract all the people names from the transcript. The extracted names are grouped by stories, since a person's name should appear in the same story as the person himself/herself.

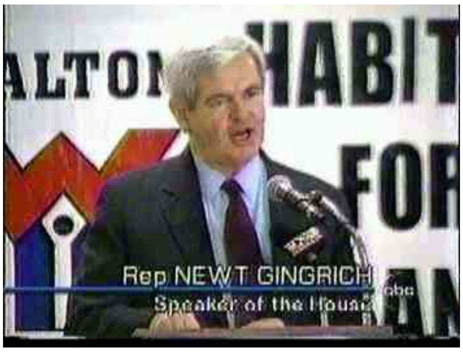
The extracted names are not useful for classifying person types until their types (anchor, reporter, or news subject) are known. This can be done precisely from the transcript clues, which imply not only the type of persons (or shots) but also the type of names. In the example clues given in Section 2.2.1, it is clear that "*Peter Jennings*" is an anchor's name while "*Linda Douglass*" and "*Barry Serafin*" are reporters' names. Since the anchors' and reporters' names are heavily recurrent in the broadcastings, they can be accurately identified by cross-verifying a large number of such clues. Once the anchors' and reporters' names are identified, the rest are all the names of news subjects. Our preliminary experiment shows that all the name types have been predicted correctly on our test dataset (see Section 4).

Many features are derived from the predicted name types (see Table 1), among which is the presence (or absence) of reporter names and subject names in a given news story. If a story does not have a reporter's name, shots in that story are unlikely to contain reporters since they rarely appear unnamed. This is almost true for subject' names, although there are occasionally anonymous subjects such as interviewees on the street. But the anchor appears virtually in every story while his/her name is rarely mentioned. More sophisticated features can be derived by examining the relations between names and speaker identities, e.g., whether the person in question utters any of the names in a story, as discussed in Section 2.2.3.

Besides the name type, the gender of each name provides another interesting clue, which is obtained by looking up the first name in the lists of common male and female (first) names. The gender of a name is set to *male* if it has a male's first name, *female* if it has a female's first name, or *both* if it can be either a male or a female's name. The gender information does not work by itself; instead, it makes sense when comparing with the estimated gender of the speech of the shot being examined. This will be further discussed in Section 2.2.3.

2.2.3. Speaker identity

The speech accompanying the news video is segmented and the segments are clustered using LIMSI speech detection and recognition engine [3]. Presumably, the speech segments in the same cluster belong to the same speaker and they



Overlaid text
Rep. NEWT GINGRICH Speaker of the House

VOCR text
rgp nev~j ginuhicij i~t thea i~ous~ i ~

Edit distance to names*:*

Newt Gingrich (0.46)
Bill Clinton (0.67)
David Ensor (0.72)
Saddam Hussein (0.78)
Bill Richardson (0.80)
Elizabeth Vargas (0.88)

Figure 4: Analysis of overlaid text by video OCR and edit distances

are assigned a unique identity (speaker ID). The gender of each speaker ID is also predicted by LIMSI. Since a shot may temporally overlap with several speaker IDs, the one with the maximal temporal coverage in the shot is regarded as the primary speaker ID of the shot. We derive the following features from the primary speaker ID of a monologue shot, including (a) whether it occurs in more than one story, (b) whether it has the largest temporal coverage in the story, (c) how many neighboring shots does it span to continuously (denoted as "span"), (d) how fast the speaker talks, (e) whether the speaker of this ID utters the anchor's, reporter's, or subjects' names, and (f) whether its gender matches with the gender of each name in the story. Figure 3 shows 6 consecutive shots with their speaker IDs. The primary speaker ID is I for shot 1 and II for the other 5 shots. The "span" of a shot is the number of consecutive adjacent shots that has the same primary speaker ID.

Feature (a) helps identify anchor shots since only the anchor's speaker ID may go across story boundaries. Feature (b) and (c) are useful for distinguishing news subject shots, because their speaker ID rarely dominates a story or spans over multiple shots. Feature (d) is calculated as the number of words uttered in a unit time, which is useful since anchors and reporters are usually faster speakers than news subjects. Feature (e) and (f) are derived from the relationships between speaker IDs and the people names detected from the transcript. Feature (e), obtained by examining the temporal overlap between speaker IDs and name occurrences, is useful since whether a person utters someone's name is informative due to news footage grammars. For example, anchors and reporters often say their own names, while news subjects rarely do; news subjects rarely say anchor's and reporter's names, while the reverse is not true. Therefore, if a person mentions the anchor's name, he/she is unlikely a news-subject. Feature (f) works in a similar way. For example, if the speaker's gender does not match the anchor's gender or the reporter's gender, he/she is probably a news-subject.

2.2.4. Video OCR (overlaid text)

Some shots of news video have short text strings overlaid on the video frames to provide key information such as locations, or names and titles of people. As shown in Figure 4, people's names appear as overlaid text in many monologue shots. Thus, if accurately recognized the overlaid text can help identify the type of a person, since the type of names are known from analyzing transcript clues. However, video optical character recognition (VOCR), the technique for recognizing overlaid text, is unlikely to produce satisfactory results on NTSC-format video which features a low resolution. Such as in Figure 4, the name "*New Gingrich*" has been recognized as "*nev~j ginuhicij*".

Though of poor quality, the VOCR text still provides weak clues pointing to the correct name of the corresponding person, and therefore, the type of the person. In the above example, the lousy VOCR text looks more similar to the correct name (*Newt Gingrich*) than other names in the transcript of the same story. Since we know *Newt Gingrich* is a news-subject, we can classify his type correctly. A similarity measure between two text strings is needed to tell how similar a name is to the VOCR text. A normalized version of *edit distance*, namely the number of insertion, removal, and substitution needed for converting one string to another, is used for this purpose. Since it is hard to tell which portion of a VOCR string corresponds to a person's name, we use a sliding window to find the portion of the VOCR string that *maximally* matches with a name in the story, and the corresponding (minimal) edit distance is used as the distance between the name and the overlaid text. The type of name with the smallest edit distance is regarded as the most likely

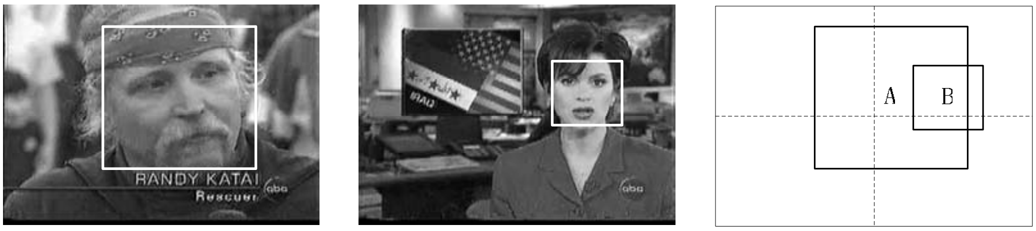


Figure 5: Two detected faces and their bounding boxes in the image frame

type of the person according to the overlaid text. Figure 4 shows the normalized edit distance from each name to the VOCR text, where the correct name (i.e., *Newt Gingrich*) has the smallest distance as expected.

2.2.5. Facial information

Each monologue shot contains a dominant face which belongs to the speaker. However, we do not rely on face recognition to predict person type for two difficulties. First, in news videos people appear in highly heterogeneous lighting and pose conditions, which will cause face recognition to fail miserably. Second, face recognition only deals with a limited set of faces stored in the database, and cannot be generalized to identify unknown faces which will inevitably emerge in new videos. Although face recognition is not applicable, the characteristics of a detected face, such as its size, location, and orientation, tell much about the *mise-en-scène* of the shot and consequently the type of person in it. For example, we find that anchors and reporters usually appear with frontal faces in the middle of the scene, and their face sizes range from small to medium. In contrast, news-subjects can have side faces, and their faces are usually larger.

To represent the location of a face, we divide a video frame into four equally-sized regions as *top-left*, *top-right*, *bottom-left*, and *bottom-right*. Faces completely falling into one of the four regions have its location feature set to this region. If a face covers both the top-left and bottom-left regions, we set its location to *left*. The *right* feature is set in a similar way. If a face does not fit into any of the aforementioned regions, its location is set to *center*. Note that we do not distinguish faces across only the top or bottom two regions since such faces are very rare. This results in totally 7 binary location features. The size feature of a face is calculated as the ratio of the area of the face's bounding box against of the frame size, with the ratio quantized into 8 discrete values. Face orientation can be "frontal", "left", or "right", denoted by three binary features. Figure 5 shows two example faces, where the one on the left is a news-subject with a large, side face in the center of the frame, while the other one is an anchor with a medium-size, frontal face on the right.

2.2.6. Temporal structure

Broadcast news footage has a relatively fixed structure. A typical news story is first briefed by an anchor, followed by several shots showing the news event, where news-subject(s) and reporter(s) appear in an interleaved manner. The story is usually, though not always, ended with the reporter or anchor giving the concluding comments. Although there are counter-examples, such as a short story consisting of only anchor shots, this structure is helpful to our task particularly for identifying anchors. We examine the position of the given shot in the sequence of shots of the corresponding news story, and use its offset (in terms of shot count) from the start and the end of the story as two structural features.

3. PERSON TYPE CLASSIFIER

The multitude and variety of features discussed in Section 2 makes the use of machine learning methods a necessity for effectively combining the features for person type classification. Support Vector Machine (SVM) [2] is a general machine learning algorithm with structural risk minimization principle. It has several nice properties that make it a suitable choice for our task. For example, we have a large number of correlated features, while SVM is capable of handling mutually dependent, high dimensional features. However, there are three class labels in our problem setting, namely anchor, reporter, news-subject, while SVM only produces a binary decision. This can be overcome by constructing "one-against-one" SVM classifiers between any two classes (anchor vs. reporter, anchor vs. news-subject, and reporter vs. news-subject), and converting the labels of a shot produced by all the classifiers into its final class through a certain "encoding" strategy (e.g., majority vote). This is done automatically by the SVM toolkit we used, which is LibSVM [6].

Table 1: The multi-modal features extracted from a shot (person) to be classified

Modality	Feature	Description
Speaker ID (*SID*)	*cross_story*	whether this *SID* appears in multiple stories
	dominant_story	whether this *SID* is the most frequent speaker ID in the story
	utter_name *(anchor, reporter, or subject)*	whether this *SID* utters the name of (any) anchor, reporter, or news-subject
	Span	number of adjacent shots this *SID* spans over continuously
	talk_speed	The number of words this *SID* utters within a unit time
	gender_match *(anchor, reporter, or subject)*	whether the gender of this *SID* matches the gender of the name of (any) anchor, reporter, or news-subject
Structure	*shot_start_offset*	order of the shot in the shot sequence from the start of story
	shot_end_offset	order of the shot in the shot sequence from the end of story
Transcript clues	*anchor_shot*	whether the transcript suggests an anchor shot
	reporter_shot	whether the transcript suggests a reporter shot
Face	*size*	the size of the face quantized into 7 discrete values
	location	the face position as center, top-left, top-right, bottom-left, bottom-right, left, right
	orientation	whether the face is a frontal, right, or left face
Video OCR	*length*	number of characters in VOCR text (if exists)
	vocr_edit_dist *(anchor, reporter, or subject)*	edit distance between the VOCR text and the name of (any) anchor, reporter, and news-subject
Named entity	*has_reporter*	whether there are any reporter's names in the transcript
	has_subject	whether there are names of any news subjects in the transcript

4. EXPERIMENTS

The test dataset used in our experiments is ABC World News Tonight video in 10 random days (30 minutes per day) in 1998, which has been also used in TREC Video Retrieval Evaluation [11]. The monologue shots to be classified are identified using a monologue detector [10] with human inspections. There are totally 498 people (or monologue shots) in the test data, among which 247 are news subjects, 186 are anchors, and the rest are reporters. The multi-modal features used in our approach are summarized in Table 1. All the features have been normalized into the range [0, 1] before being fed into the SVM classifier.

A 10-fold cross-validation is used to evaluate the classification performance. Each time we train the classifier using the 9 out of the 10 days' videos and test it on the news video of the remaining day. This is repeated 10 times, each time using a different day's video for testing, and the performance measures are averaged over the 10 runs. The performance is evaluated using precision and recall on each type of person, defined as:

$$Precision = \frac{|C \cap C'|}{|C'|} \quad \text{and} \quad Recall = \frac{|C \cap C'|}{|C|}$$

where C is the set of persons of a certain type, and C' is the set of persons that are classified as this type by our method. The performance of the 10-fold cross validation is shown in Table 2.

Table 2: Performance of person type classification

	Overall (498)	Anchor (186)	Reporter (65)	News Subject (247)
Correct	466	180	51	235
Miss	32	6	14	12
False Alarm	32	4	10	18
Precision	0.936	0.978	0.836	0.929
Recall	0.936	0.968	0.785	0.951

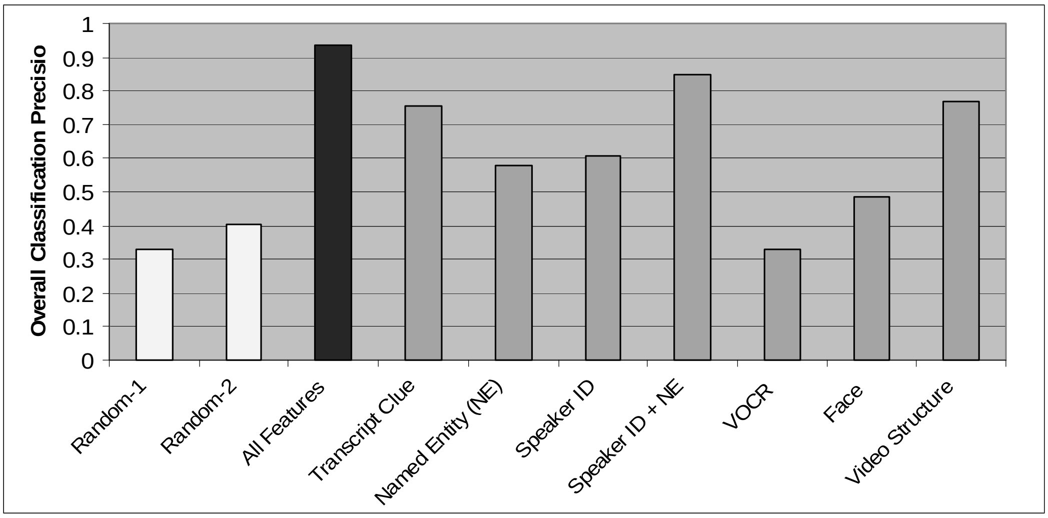


Figure 6: Contribution of different categories of features

As we can see, our classifier can almost perfectly identify anchors, resulting in less than 10 misses and false alarms. Its performance on classifying reporters and news-subjects is reasonably good, though not as high as that on anchors. A closer examination reveals that most of the misses on reporters are false alarms on news-subjects, and vice versa, which indicates that our classifier sometimes fails to discriminate reporters from news-subjects. The overall precision and recall are equivalent (since the misses of one class are false alarms on other classes), which is 93.6%. Note that a random 3-class classifier can achieve 33% precision (denoted as Random-1), and a random classifier taking into account the frequencies of the three types of people can achieve 40.2% precision (Random-2). Given these two baselines, our classifier is very effective in classifying person types.

In addition, we studied the contribution of different categories of features in our classifier. For this purpose, we re-train the classifier based on the features of each category and test its performance using 10-fold cross-validation. The overall classification precision achieved by each category of features are plot together with that achieved by using all the features as well as that of the two random classifiers in Figure 6. As shown, speaker identification, video structure, and transcript clues are the most effective categories of features, which alone achieves around 70% overall precision. In comparison, video OCR and face information are quite useless as they do not significantly outperform the random classifiers. Moreover, as discussed in Section 2, speaker identification and overlaid text will result in more effective features if combined with named entities. This is apparent from Figure 6. Although named entity features are not very effective by themselves, when combined with speaker ID features the precision rises from 61% (speaker ID) to 86% (speaker ID + named entity). Overall, this study demonstrates the benefit of multi-modal analysis, especially the relatively under-studied audio/speech features, as well as the importance of combining features from different modalities.

5. CONCLUSIONS

We have described a classifier for discriminating person types in broadcast news video based on multi-modal analysis, which has been proved effective on TRECVID dataset. This work gives a typical example on how to analyze different video modalities including speech, transcript text, video frames to derive features useful for a specific high-level video analysis task, and how to combine the multi-modal features with a learning approach. Though the features used in this work are task-dependent, the general framework on multi-modal analysis is applicable to many other video analysis tasks. Future works on this direction include labeling the persons appearing in news video with names and roles.

REFERENCES

1. Bikel, D. M., Miller, S., Schwartz, R., and Weischedel, R., Nymble: a high-performance learning name-finder. In Proc. 5th Conf. on Applied Natural Language Processing, 1997, pp. 194-201.

2. Burges, C. J. C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2):121-167, 1998.

3. Gauvain, J.L., Lamel, L., and Adda, G. The LIMSI broadcast news transcription system. Speech Communication, 37(1-2): 89-108, 2002.

4. Houghton, R. Named Faces: Putting Names to Faces. In IEEE Intelligent Systems Magazine, 14(5): 45-50, 1999.

5. Jin, R., Hauptmann, A. Learning to identify video shots with people based on face detection. In IEEE International Conference on Multimedia & Expo, Baltimore, MD, USA, July 6-9, 2003.

6. LIBSVM. A library for support vector machine. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

7. Sato, T., Kanade, T., Hughes, E. K., Smith, M. A., Satoh, S. Video OCR: Indexing digital news libraries by recognition of superimposed caption. ACM Multimedia Systems, 7(5): 385-395, 1999.

8. Satoh, S., Y., Kanade, T. NAME-IT: Association of Faces and Names in Video. In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, 1997, pp. 368-373.

9. Schneiderman, H., Kanade, T. Object detection using the statistics of parts. Int'l Journal of Computer Vision, 56(3): 151-177, 2002.

10. Snoek, C.G.M. and Hauptmann, A. *Learning to identify TV news monologues by style and context.* Technical Report, CMU-CS-03-193, Carnegie Mellon University, 2003.

11. TRECVID: TREC Video Retrieval Evaluation: http://www-nlpir.nist.gov/projects/trecvid/.

12. Yang, J., Chen, M, Hauptmann, A. Finding Person X: Correlating Names with Visual Appearances, Int'l Conf. on Image and Video Retrieval, Dublin City, July 21-23, 2004. (to appear).

13. Yang, J., Hauptmann, A. Naming every individual in news video monologues. ACM Multimedia 2004, New York City, Oct. 10-16, 2004..

14. Zhang, H.J., Tan, S.Y., Smoliar, S.W., Gong, Y.H. Automatic parsing and indexing of news video. In Multimedia Systems, 2(6): 256-266, 1995.