Anchor Shot Detection in TRECVID-2005 Broadcast News Videos

Akira Yanagawa, Winston Hsu, and Shih-Fu Chang

Dept. of Electrical Engineering, Columbia University, New York
{akira, winston, sfchang@ee.columbia.edu}

Columbia University ADVENT Technical Report #213-2005-7

ABSTRACT

In this paper, we discuss a new method for detecting anchor shots in broadcast news video. Our approach makes use of face information that includes the position, the size and the number of faces detected in the image frame. To alleviate the adverse effects caused by occasional face detection errors, we propose two new ideas based on multiple queries and zero face padding. Our experiments over TRECVID 2005 data confirm the robustness of our approach – 97.6% precision at 85.7% recall and improves the baseline up to the maximum of about 15% in recall with the almost same precision.

1. INTRODUCTION

Due to the explosion of broadcast channels, news stories, also a major source for information exploitation, increase exponentially. An efficient to organize these large-scale video databases is required. Especially, in video indexing and searching, to reinforce the accuracy of retrieving images, pre-processing procedures to remove non-informative video segments (e.g., anchor shots, commercials) is necessary. Among them, anchor shot detection is the most important in news video databases [1, 3] because anchor shots usually carry no useful information but anchorpersons in studios.

There have been several prior works in anchor shot detection. Some of them made use of face information such as the position, the size, and the number of faces. Winston Hsu, Shih.-Fu Chang , et al. extracted face regions and by utilizing the color histogram within the face region and temporal coherency of the face size and position to detected anchor shots [2]. Seung-Chul Jun and Sung Han Park also utilize face information to detect anchorperson [6]. Besides, Z. Liu and Q. Huang made use of anchorperson's attire for anchor detection [7]. Xiaodan Song, Ching-Yung Lin et al., applied face recognition methodology for anchor shot detection[12]. All these conventional methods heavily rely on face detection results. Therefore, the performance is bounded by

face detection accuracy. If the face detection result is poor, it is difficult for these anchor detection algorithms to keep fair performance since face detection is a strong cue. According to our experiment, applying OpenCV [4] to detect faces in TRECVID2005 dataset [8], the face detection algorithm incurs certain errors including false positive and miss.

Besides, HongJiang Zhang, et al. has completed pattern matching with several spatial structures of anchorperson shots that were classified before hand [13], where they showed that the anchor shots retain coherence not only in the anchor face regions but also in the backgrounds.

In this paper, we proposed novel approaches to take advantage of the face detection information and also try to tolerate certain face detection errors by utilizing the coherence of special structures of anchorperson shots. Among them, multi-query is to deal with the situations of multiple detected faces within the image or false alarms caused by face detection algorithms by using noisy-or fusion method; zero-face-padding is to relieve the problem of miss in face detection step. We conduct the experiments in TRECVID 2005 dataset and confirm the robustness of our approach – 97.6% precision at 85.7% recall and improves the baseline up to the maximum of about 15% in recall with the almost same precision.

We defined an anchor shot as the image that includes only anchorpersons in a studio. We therefore don't have to know who the anchor is but the presence of the anchorpersons. With according features, we classified the image as an anchor shot or not. We chose SVM with RBF kernel as classifier because this classifier showed good performance for image retrieval [1].

In section 2, we discuss low-level features that were selected for anchor shot detection. In section 3, we explain our novel method to omit the efficiency of face detection's accuracy.

2. LOW LEVEL FEATURES

To detect anchor shot, we propose to use several image features rather than pixel-level intensity data such that we can reduce computational complexity as well as benefit from effective feature representations.

As we discussed in the introduction, face information is effective for anchor shot detection, and an entire image is also useful because it has some coherency in the same channel. Because of that, we utilize not only local features related with face regions but also global features of the entire images for anchor shot detection.

Specifically, we selected color histogram and grid color moment for global features of the whole image, and grid color moment on small face regions. We describe the features in the next sub-sections.

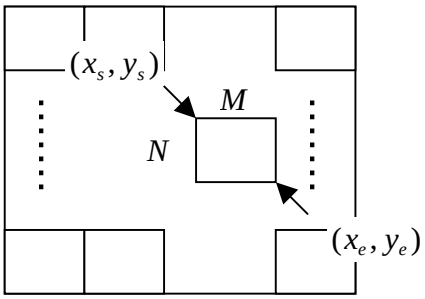

Fig. 2-2-1 Color Moment

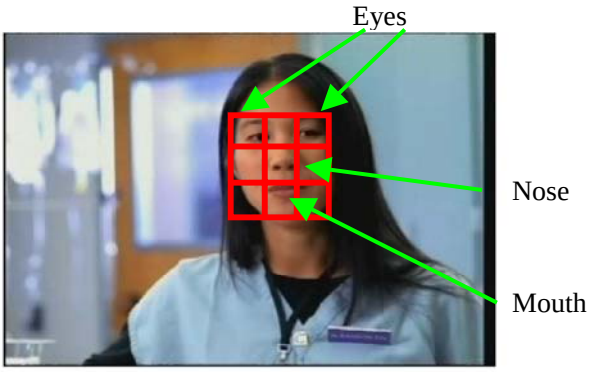

Fig. 2-2-2 Grid Color Moment on Face region

2.1. COLOR HISTOGRAM

Color histograms describe the color distribution in an image. Color histograms have the advantage that this is invariant to affine transformation though it has the disadvantage that it can't retain local relationship. Since there are many image variations even in the same shot due to camera pan or tilt, the invariant to affine transform is more important to denote a global feature. Because of that, we chose color histogram as one of image features.

In practice, we use color histograms that have 166 bins in HSV color space [9, 10]. This color histogram can reduce the number of colors drastically. In addition, by focusing on hue and taking gray (no hue) in account, this color histogram can describe color distribution well though the number of bins is decreased from 2^{24} to 166.

2.2. GRID COLOR MOMENT

Color histograms are useful in describing the color distribution in an image. However, since there is no local relationship between information in color histograms, it can't sometimes make differentiations between objects. To compensate for this, we utilized grid color moment in addition to color histograms.

Color moment denotes the color distribution by using mean, standard deviation and third root of the skewness of each color channel [11]. If the value of the ith color channel at the (x,y) image pixel is , and the width and height of grid are M and N respectively then the color moments are:

$$E^i = \frac{1}{MN}\sum_{y=y_s}^{y_e}\sum_{x=x_s}^{x_e} p_{x,y}^i \quad \sigma^i = \left[\frac{1}{MN}\sum_{y=y_s}^{y_e}\sum_{x=x_s}^{x_e}\left(p_{x,y}^i - E^i\right)^2\right]^{\frac{1}{2}} \quad s^i = \left[\frac{1}{MN}\sum_{y=y_s}^{y_e}\sum_{x=x_s}^{x_e}\left(p_{x,y}^i - E^i\right)^3\right]^{\frac{1}{3}}$$

Where x_s, x_e, y_s *and* y_e are the start and end x coordinate and the start and end y coordinate respectively (Fig. 2-2-1).

In anchor shot detection, for a global image feature, an image is divided 5 by 5 grids with calculated color

moment per grid. In addition, we utilized grid color moment for local feature because the color moment of each grid has the local relationship information that denote the each color distribution of face parts like nose, mouse and eyes (Fig. 2-2-2). For color space, we used LUV which is suitable for image recognition [5].

2.3. FACE DETECTION

We used Open CV [4] to detect faces. From the result of face detection, we utilize the number of faces, the center position of each face and area of each face in an image as featured.

This face detection methodology is a much less computational effort than other conventional methods, and results in a low false positive rate and a high detection rate. However, there are some false alarms and misses in news video images because in general, video image is low resolution and noisy compared to still images. Moreover, unfortunately this methodology cannot detect profile and small face (Fig. 2-3-1).

Fig. 2-3-1 False alarm and miss

3. SOLUTIONS FOR ACCURACY

Since anchorpersons are always in anchor shots, face information is effective in anchor shot detection; however, the accuracy rate of that makes the upper bound of anchor shot detection.

If face detector reports non-face region as face region (false alarm), causes the result of anchor shot detection degraded. Although grid color moment on face region and other global image features support to increase precision ratio, it does not work enough for recall ratio, if the number of false alarms is large.

In case of a miss, it is possible that more serious damages will be occurred. In anchor shot detection, since we assume that faces must be in anchor shot, if face detector reports there is no face in an image, the image will not be processed anymore. It decreases recall ratio.

To resolve these two problems, we suggested multi-query model and zero face padding. Multi-query model reduces the effect of false alarm. Likewise, zero face padding reduces the effect of miss.

In the next section, each methodology is described in detail.

3.1. MULTI-QUERY

To reduce the effect of false alarms coming from the face detector, we introduced noisy-or like model. In noisy-or model, if at least one parent cause is present, the result will be true. In our system, if at least one candidate of face is detected as anchor shot, this image will be anchor shot even though there are some candidates that are detected as non-anchor shot in the same image. Because of that, recall ratio in our system is hardly affected by false alarms from the face detector.

To implement noisy-or like model on our system, we utilized multi-query (See Fig. 3-1-1). At first, we divide an image into the number of the face regions. Then each image is posted as one of multi queries to the SVM classifier for calculating the confidence whether the image is anchor shot or not. After calculated the confidence of all the queries in the image, maximum confidence of anchor shot is regarded as the result of the image.

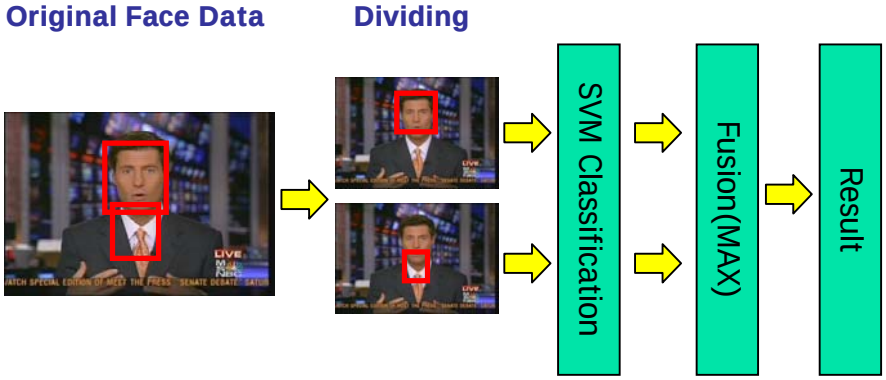

Fig. 3-1-1 Multi-Query Flowchart

3.2. ZERO FACE PADDING

Compared to handling the false alarms of face detection, handling the miss is more difficult. As far as we utilize the single face detector, we have no way to correct the result. Even though we use several face detectors, it would be hard to select the correct one from some candidates. Alternatively, it is possible to solve this problem by developing new face detectors; however, it is not realistic to eliminate all errors. Therefore, it is practical to use other information rather than face other data for compensating for the miss information.

As we discussed in the introduction, anchor shot has some coherency not only in anchor face region but also in background. By taking advantage of this property, we intentionally padded zero face images to compensate miss information from the face detector (See Fig 3-2-1). In other words, by padding the image without face information, our system can detect anchor shot with independent of the face information. Especially in case that anchorperson's face is so small that the face detector can't find it, this methodology is effective.

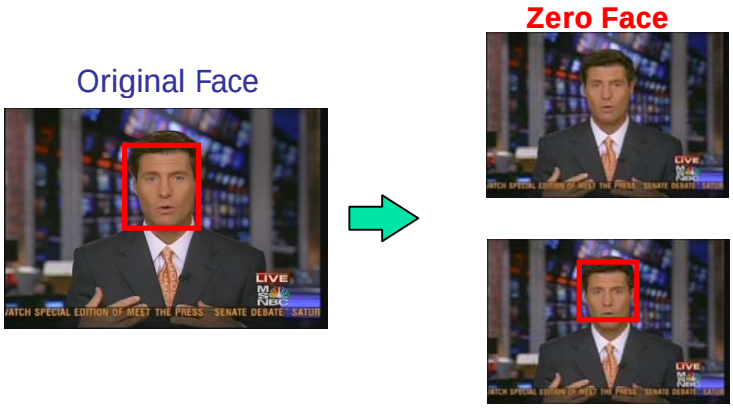

Fig. 3-2-1 Zero Face Padding

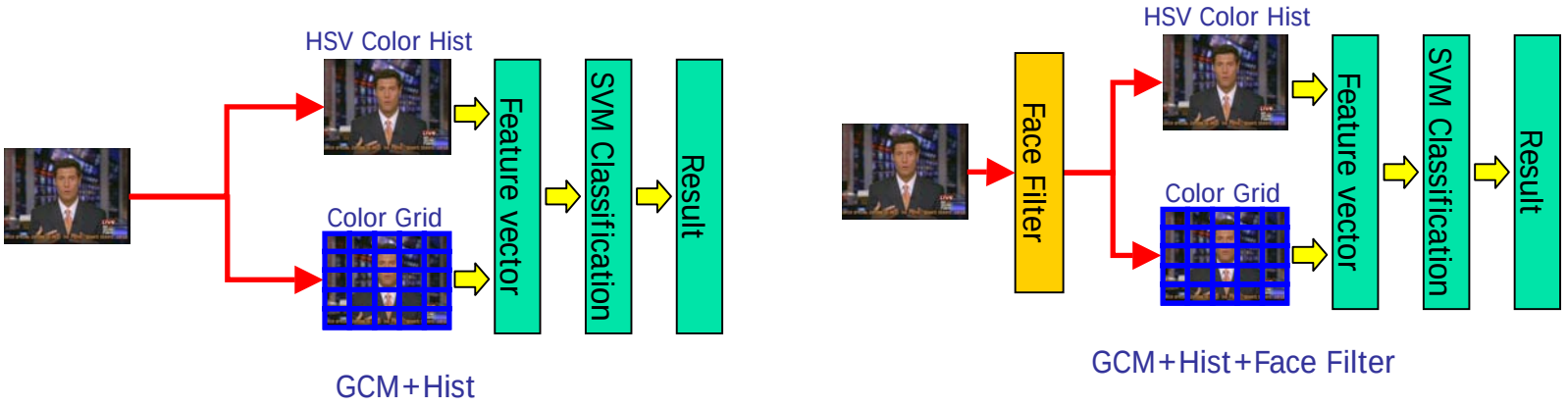

Fig. 4-1 Two Baseline methods

4. Experiments and Results

We tested to verify our anchor shot detection on TRECVID2005 video news dataset. In the dataset, there are six TV stations: CCTV (China), CNN (U.S.), and LBC (Lebanon), MSNBC (U.S.), NBC (U.S.), NTDTV (U.S.-based Chinese-language). We divided the dataset into training and test sets with respect to each TV station. In training set, after extracted image features, we trained SVM (RBF kernel) by using k-fold cross-validation.

In order to represent, in the entire dataset, we corrected the face region reported from the face detector by hand. We regarded these corrected dataset as ground truth. In our experiment, we compared the recall and precision ratio (Fig. 4-2) of our methods and the result of baseline methods (Fig. 4-1). One baseline methods utilized only Grid Color Moment (GCM) and Color Histograms (Hist). The other one used the same features; however by using the result of face detection, if there is no face in an image, we consder the image as non anchor shot immediately. In addition, to study our methodology in detail, we calculated the average precision of these three method (Fig. 4-3).

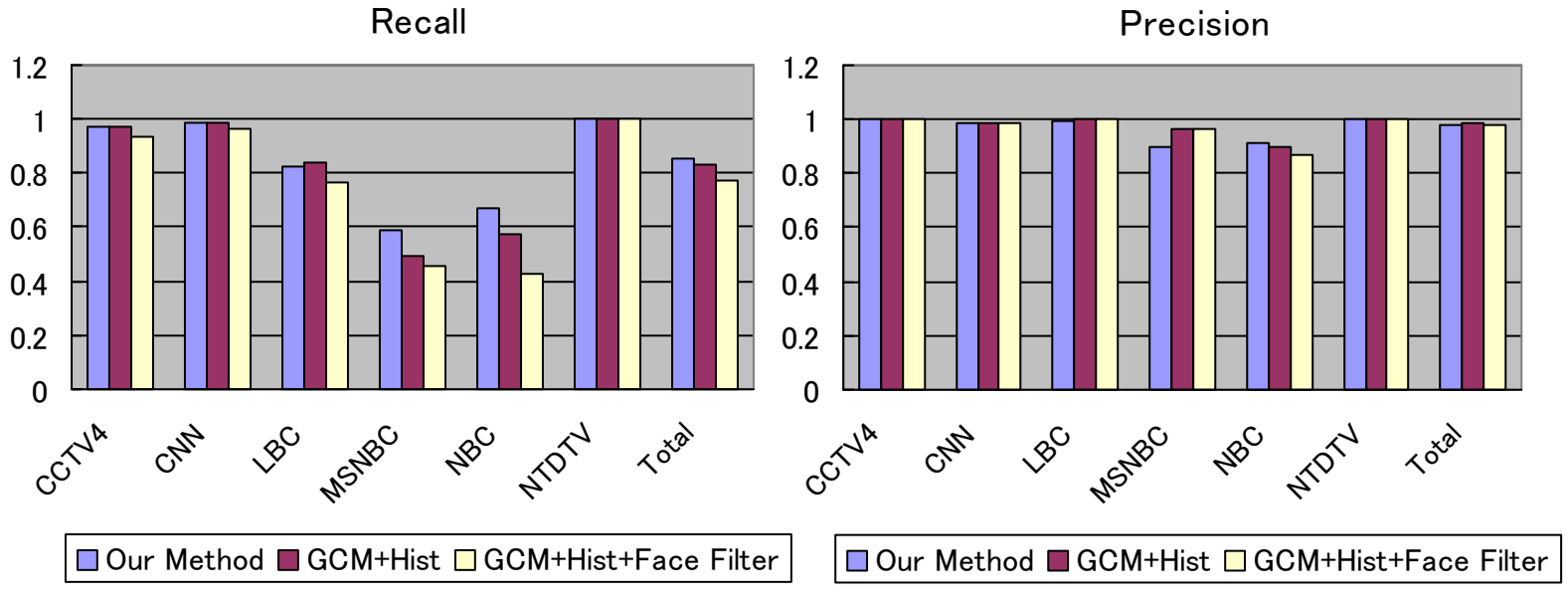

Fig. 4-2 Recall and Precision

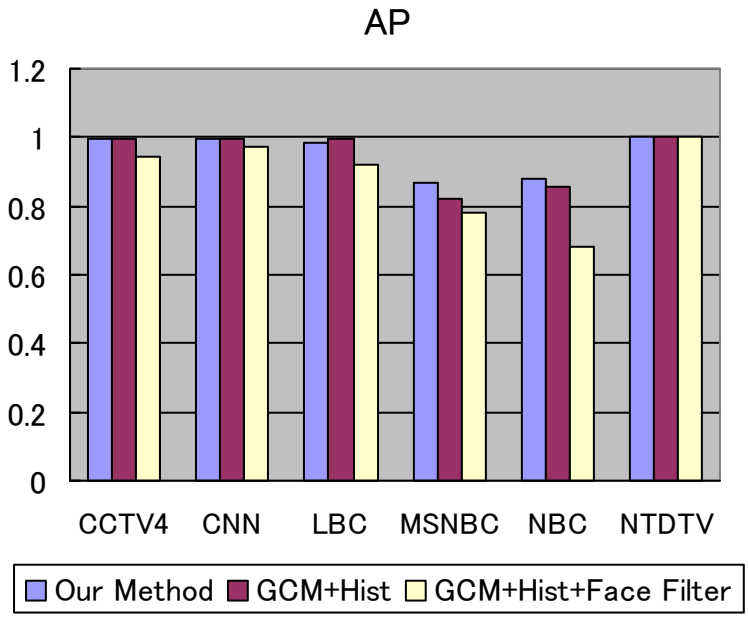

Fig. 4-3 Average Precision

5. Conclusions and Observations

Comparing the results of our method and the baseline methods with respect to recall and precision, the recall of our method is as good as or better than that of the baseline methods with high precision. In MSNBC and NBC, our methods improved recall up to about 15%. In general, the accuracy is limited by pre-process. This result showed our anchor shot detection is hardly restricted by the face detection (pre-process). As the result of that, the average precision (AP) of our method is quite good over the TV stations. All of the AP exceeded 0.85 and three TV stations' APs out of five are nearly equal one.

The results of MSNBC and NBC are not as good as the other TV stations. The reason is that comparing them to the news programs of the non-U.S. TV stations, in the news programs of U.S. TV stations, an anchor shot tends to be complicated. Because of that, spatial structure anchorperson shot didn't work efficiently in MSNBC and NBC. However, since our method makes use of face information efficiently, the performance of our method exceeded to the results of the baseline methods.

Finally, in this paper, we focused on the influence from false face information rather than classifier or features. Although we could get the result reasonably well in our experiment, to make further refinements, we will study classifier and features suitable for anchor shot detectors.

6. Acknowledgement

We thank Lexing Xie, Dongqing Zhang, Eric Zavesky and Lyndon Kennedy of Columbia University for their useful advice. This material is based on work funded in whole by the U.S. Government. Opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do reflect the views of

the U.S. Government.

7. References

[1] A. Amir, J. O Argillander, M. Berg, S.-F. Chang, M. Franz, W. Hsu, G. Iyengar, J. R Kender, L. Kennedy, C.-Y. Lin, M. Naphade, A. (Paul) Natsev, J. R. Smith, J. Tesic, G. Wu, R. Yan and D. Zhang, *IBM Research TRECVID-2004 Video Retrieval System*, *NIST TRECVID 2004 Workshop*, Gaithersburg, MD, 2004.

[2] W. Hsu, S.-F. Chang, C.-W. Huang, L. Kennedy, C.-Y. Lin and G. Iyengar, *Discovery and Fusion of Salient Multi-modal Features towards News Story Segmentation*, *IS&T/SPIE Symposium on Electronic Imaging: Science and Technology - SPIE Storage and Retrieval of Image/Video Database*, San Jose, USA, 2004.

[3] W. Hsu, L. Kennedy, S.-F. Chang, M. Franz and J. Smith, *Columbia-IBM News Video Story Segmentation In TRECVID 2004*, *Columbia ADVENT Technical Report*, 2004.

[4] Intel, *Compute-Intensive, Highly Parallel Applications and Uses*, Intel Technology Journal, 09 (2005).

[5] X. Y. Jin. and e. al, *Feature evaluation and optimizing feature combination for content-based image retrieval*.

[6] S.-C. Jun and S. H. Park, *An Automatic News Video Semantic Parsing Algorithm*, *ITC-CSCC2001*, 2001.

[7] Z. Liu and Q. Huang, *Adaptive Anchor Detection Using On-line Trained Audio/Visual Model*,, *Proc. of SPIE*, 2000.

[8] NIST, *TREC video retrieval evaluation (TRECVID)*, 2001-2005.

[9] J. R. Smith and S.-F. Chang, *Tools and Techniques for Color Image Retrieval*, *IS&T/SPIE Symposium on Electronic Imaging: Science and Technology (EI'96)*, 1996.

[10] J. R. Smith and S.-F. Chang, *VisualSEEk: a Fully Automated Content-Based Image Query System*, *ACM Multimedia*, 1996.

[11] M. Stricker and M. Orengo, *Similarity of color images*, *in Storage and Retrieval for Image and Video Databases (SPIE)*, 1995.

[12] Xiaodan Song, C.-Y. Lin and M.-T. Sun, *Cross-Modality Automatic Face Model Training from Large Video Databases*, *The 1st IEEE Workshop on Face Processing in Video*, 2004.

[13] H. Zhang, Y. Gong, S. W. Smoliar and S. Y. Tan, *Automatic Parsing of News Video*, *Proc. IEEE Conf. on Multimedia Computing and Systems*, 1994.