ANALYSIS OF VISUAL SIMILARITY IN NEWS VIDEOS WITH ROBUST AND MEMORY-EFFICIENT IMAGE RETRIEVAL

D. M. Chen, P. Vajda, S. S. Tsai, M. Daneshi, M. C. Yu, H. Chen, A. F. Araujo, B. Girod

Department of Electrical Engineering, Stanford University, Stanford, CA, USA
{dmchen, pvajda, sstsai, mdaneshi, mattcyu, hchen2, afaraujo, bgirod}@stanford.edu

ABSTRACT

Many large collections of news videos dating back several decades can now be accessed online. For users to easily retrieve a compilation of stories on a particular event/topic and to quickly sample each story clip, all the news videos must be precisely segmented into stories and a representative video summary must be generated for each story. In this paper, we demonstrate that effectively exploiting the visual similarities pervasive in all news videos can greatly help to fulfill these technical requirements and thus enable the dynamic retrieval and mixing of small news video fragments. Two new algorithms are developed to accurately detect two important sources of visual similarity: (1) similar preview and story frames, and (2) repeated appearances of a news anchor. As a result, valuable sources of preview clips and informative clues about story boundaries are obtained from identification of these visual similarities. The retrieval engine implemented in both algorithms employs compact global image signatures and requires a small memory footprint, so that many instances of the detection algorithms can run concurrently on the same server for fast processing of a large collection of news videos. At the same time, the retrieval engine is robust to the large appearance variations encountered in the preview matching and anchor detection problems. In addition, since the video frame's color information is not required in our algorithms, both modern color and vintage black-and-white news footage can be processed in the same framework.

Index Terms— Visual similarity, news video analysis, story segmentation, preview detection

1. INTRODUCTION

Large collections of news videos now exist at many universities, institutions, and companies. Examples include the Vanderbilt Television News Archive[1], Internet Archive[2], and TVEyes[3]. These large collections are immensely valuable to journalists, historians, and researchers as well as to general consumers. Finding the exact video fragments which correspond to stories of a particular event or topic within these large archives, however, is still challenging. It is highly desirable to automatically (1) generate a compilation of accurately segmented news clips for a user-specified query and (2) provide short video summaries so that customers can preview each clip before deciding to borrow or purchase the entire clip. To achieve these two goals, each episode of a television news program must be precisely segmented into individual stories, and a short but highly representative video summary must be generated for each story.

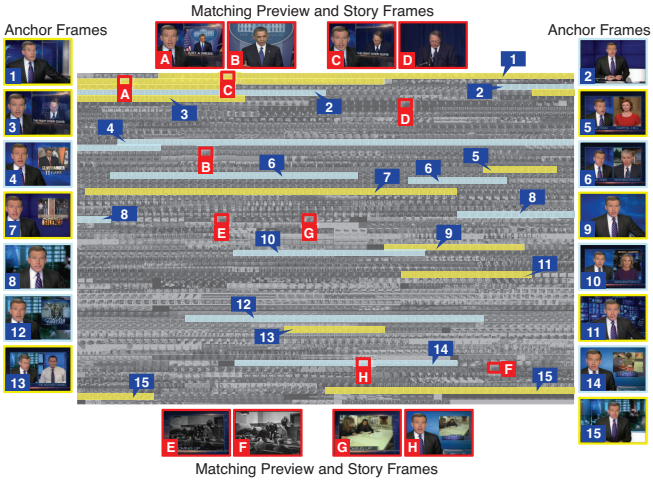
In this paper, we address the aforementioned technical challenges by exploiting visual similarity within news videos. Visual similarities are found in abundance throughout a news broadcast and reveal valuable information about the structure of the broadcast. Fig. 1(a) and (b) show mosaics of keyframes for a 30-minute episode of NBC Nightly News and a 30-minute episode of ABC World News. Two important sources of visual similarity can be identified in the figure:

- Frames in teasers or previews are visually similar to frames in the actual stories that appear later. A few examples of matching preview and story frames are highlighted with red borders and marked with letters A, B, C, D, ...
- Frames in an anchor shot are visually similar to frames in another anchor shot. Keyframes from all anchor shots are highlighted with yellow and cyan borders and marked with numbers 1, 2, 3, 4, ...

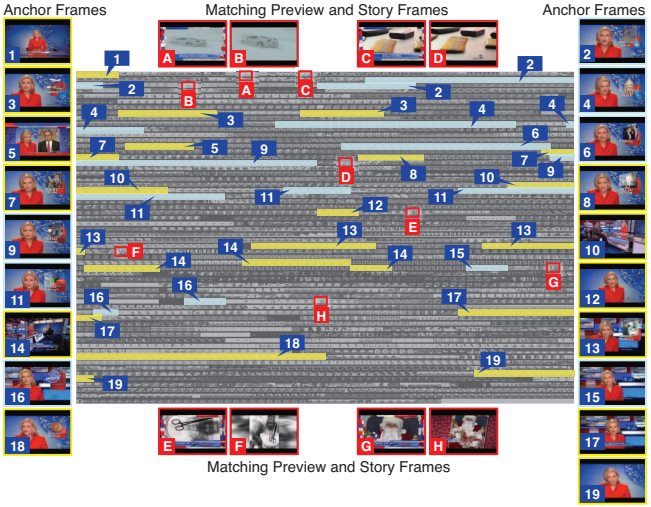
Accurate detection of these visual similarities is useful for analysis, organization, and mixing. First, by matching the preview frames to corresponding story frames, we obtain (1) a rich source of preview clips and (2) informative clues about story boundaries because a preview typically precedes a story or a commercial break. Second, by detecting appearances of a news anchor, we obtain (1) another source of preview clips because many stories start with introductory statements by the anchor, (2) the option of removing the anchor shots and focusing on just the contributed stories, and (3) important cues for story segmentation because anchor appearances are highly correlated with story boundaries.

Anchor appearance has been frequently used as a higher-level visual feature in many story segmentation methods; a

[1] http://tvnews.vanderbilt.edu
[2] http://archive.org/details/tv
[3] http://www.tveyes.com/



(a) NBC Nightly News on December 21, 2012



(b) ABC World News on December 20, 2012

Fig. 1. Gray blocks in the center are mosaics of keyframes. Visually similar preview/story frames or anchor frames are highlighted in red or yellow/cyan, respectively.

good review of the topic is available in [1]. Zhang et al. [2], Hanjalic et al. [3], and Liu and Huang [4] have developed model-based and template-based anchor detection algorithms. Alternatively, Gao and Tang [5], De Santo et al. [6], Gao et al. [7], D'Anna et al. [8], Ma and Lee [9], and Broilo et al. [10] perform model-free anchor detection by clustering of frames in different feature spaces or by graph theoretical analysis on shots. Liu et al. [11] and Zheng et al. [12] employ spatio-temporal slices for visual analysis. Anchor detection has also been addressed from a classification perspective by Bertini et al. [13], Xiao et al. [14], and Lee et al. [15].

Robust detection of visual similarity in news videos is challenging because there are various distortions between similar frames including image scale changes, brightness/contrast deviations, object movement, cropping, background variations, and clutter, as can be seen from Fig. 1. Additionally, efficient detection methods are required to process a large set of news videos. We design and develop preview matching and anchor detection algorithms which are resilient against many types of visual distortions. For preview matching, our algorithm first detects preview frames using on-screen text features and then retrieves visually similar story frames using local image features. For anchor detection, our algorithm retrieves similar frames for every frame, identifies frames which have multiple matches throughout the video, and removes false positive detections by comparing the candidate anchor frames to each other. The image retrieval engine used in both algorithms achieves high retrieval performance in spite of large geometric and photometric variations. At the same time, since the retrieval engine is extremely memory-efficient, more instances of these algorithms can run concurrently on a multi-core server for a given system memory limit, making our detection algorithms attractive for analyzing large news video collections.

The algorithms presented in this paper will only utilize the grayscale information in video frames, and therefore the algorithms can be applied to both modern color and vintage black-and-white news clips, which is important for analyzing historical news footage in archives. In contrast, the vast majority of current anchor detection algorithms require color information of the anchor, studio background, and on-screen graphics to detect human faces and to generate color histograms or moments for visual similarity comparison. Reliance on color information constrains these algorithms to process only color news clips.

The rest of the paper is outlined as follows. Sec. 2 presents our algorithm for robustly matching preview frames to story frames based on a memory-efficient retrieval engine. Then, Sec. 3 follows by describing our algorithm for anchor detection, which utilizes the same retrieval engine. Experimental results in Sec. 4 demonstrate the effectiveness of our algorithms on a heteregeneous set of news videos.

2. ALGORITHM FOR PREVIEW MATCHING

Visually similar preview and story frames can still differ considerably by geometric and photometric transformations as well as adjacent clutter such as people and computer-generated graphics. We have developed a preview matching pipeline as shown in Fig. 2 for robustly and efficiently matching preview and story frames in spite of these visual distortions. The first two blocks in the pipeline detect preview frames and define the preview regions of interest, while the subsequent four blocks retrieve similar story frames for the detected preview frames. Both the text and local image features in the video frames are utilized for accurate preview frame detection, and a database is constructed for storing the global signatures computed from the local image features of video frames.

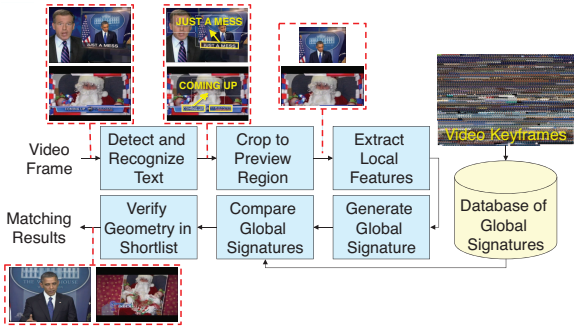


Fig. 2. Proposed preview matching pipeline.

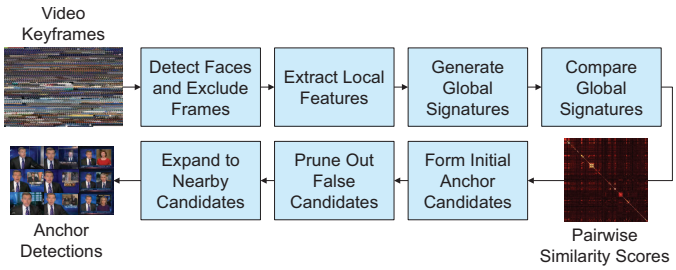


Fig. 3. Proposed anchor detection pipeline.

2.1. Preview Frame Detection

The first step in the pipeline of Fig. 2 is detection of text in a video frame with a detector based on edge-pruned maximally stable extremal regions [16] and recognition of the detected text with the Tesseract OCR engine [17]. We consider two types of preview frames: Type A frames occur at the beginning of the broadcast and Type B frames occur later in the broadcast prior to commercial breaks. Type A frames are detected as those frames which precede the program's opening logo transition and contain at least one text box. In contrast, Type B frames are detected as those frames that contain at least one text box and have a news-specific transition phrase such as "Coming Up" recognized by OCR in one of the text boxes.

The next processing step is an adaptive cropping of the preview frame to the preview region that is designed to greatly reduce false positive image matches during retrieval. In Fig. 2, it can be seen that the preview region is usually adjacent to the anchor or to a computer-generated banner. If the entire frame is matched against a database of frames from the rest of the video, there will be false positive matches to other frames showing the anchor or the same type of banner. Thus, cropping out the interfering portions of the frame and retaining just the preview region can significantly improve matching accuracy. For Type A preview frames, the region above and right of the largest detected text box is kept. For Type B preview frames, the region above the largest detected text box is kept as the matching target. In both cases, the adaptive cropping is performed automatically.

2.2. Story Frame Retrieval

After adaptive cropping of the preview frame to the preview region, the subsequent steps in the pipeline of Fig. 2 focus on retrieving similar story frames. First, local features such as SURF [18] are extracted from the grayscale version of the cropped preview region. Then, the local features are aggregated using the Residual Enhanced Visual Vector (REVV) [19] into a memory-efficient global image signature to summarize the most relevant image characteristics. REVV signatures are binary vectors that can be compared directly in the compressed domain using bitwise XOR and can be stored compactly in a small amount of memory. The REVV signature of the preview frame is quickly compared against REVV signatures of keyframes throughout the news video to generate a ranked list of the most similar story frames. Temporal nonmaximum suppression is applied within the ranked list to identify database candidates which are separated in time by 1 second or more. Finally, the story frames within the shortlist are verified with RANSAC [20] for geometric consistency in matching local feature keypoint locations with respect to the preview frame.

Since the REVV-based retrieval engine is memory-efficient, it is feasible to run different instances of the preview matching algorithm for many news videos concurrently on a multi-core server. This is especially important for analyzing a large collection of news videos. Incremental database updates can also be performed efficiently because the REVV signatures for different database frames are encoded independently. At the same time, REVV signatures are discriminative for large-scale retrieval, so the algorithm can quickly and reliably match preview and story frames in spite of considerable visual distortions. Experimental results in Sec. 4 will demonstrate high matching accuracy across many different videos with low system memory usage.

3. ALGORITHM FOR ANCHOR DETECTION

One distinctive property of the news anchor is that he/she reappears frequently throughout an episode. This property can be seen from the mosaics in Fig. 1 and can also be captured accurately by computing REVV similarity scores between every pair of keyframes. We now present an effective anchor detector that first exploits intra-episode visual similarity and then further exploits inter-episode visual similarity.

3.1. Intra-Episode Detection

The steps of the intra-episode anchor detection pipeline are depicted in Fig. 3. First, human faces are detected in the grayscale version of video keyframes using the Viola-Jones detector [21], and the frames that contain no faces are ex-

Algorithm 1 Intra-episode anchor detection.

Require: Function $S(F, G)$ which evaluates the similarity between global image signatures for frames F and G.

Require: Function $R(F, G)$ which evaluates the number of feature matches between frames F and G using RANSAC.

Require: Function TNS $(\mathbf{F})$ which performs temporal nonmaximum suppression on a set of frames $\mathbf{F}$.

Require: Timestamp $T(F_i)$ for frame F_i, where $1 \leq i \leq N_{\text{frames}}$ and N_{frames} is the number of keyframes.

Formation of Initial Candidates

$\mathbf{F}_{\text{initial}} = \emptyset$
for $i = 1 \rightarrow N_{\text{frames}}$ **do**
  $\mathbf{F}_{\text{similar}} = \{F_j : S(F_i, F_j) > T_{\text{score}}\}_{j=1}^{N_{\text{frames}}}$
  $\mathbf{F}_{\text{distant}} = \{F_j : |T(F_i) - T(F_j)| > T_{\text{time,high}}\}_{j=1}^{N_{\text{frames}}}$
  $\mathbf{F}_{\text{neighbors}} = \mathbf{F}_{\text{similar}} \cap \mathbf{F}_{\text{distant}}$
  $\mathbf{F}_{\text{neighbors,max}} = \text{TNS}(\mathbf{F}_{\text{neighbors}})$
  $\mathbf{F}_{\text{ransac}} = \{F \in \mathbf{F}_{\text{neighbors,max}} : R(F, F_i) > T_{\text{ransac,low}}\}$
  if $|\mathbf{F}_{\text{ransac}}| > N_{\text{ransac}}$ **then**
    $\mathbf{F}_{\text{initial}} \coloneqq \mathbf{F}_{\text{initial}} \cup F_i$
  end if
end for

Pruning of False Candidates

$\mathbf{F}_{\text{pruned}} = \emptyset$
for $F \in \mathbf{F}_{\text{initial}}$ **do**
  $\mathbf{F}_{\text{distant}} = \{G \in \mathbf{F}_{\text{initial}} : |T(F) - T(G)| > T_{\text{time,high}}\}$
  $\mathbf{F}_{\text{distant,max}} = \text{TNS}(\mathbf{F}_{\text{distant}})$
  $R_{\text{distant}} = \text{mean}(\{R(F, G) : G \in \mathbf{F}_{\text{distant,max}}\})$
  if $R_{\text{distant}} > T_{\text{ransac,low}}$ **then**
    $\mathbf{F}_{\text{pruned}} \coloneqq \mathbf{F}_{\text{pruned}} \cup F$
  end if
end for

Expansion to Nearby Candidates

$\mathbf{F}_{\text{intra}} = \mathbf{F}_{\text{pruned}}$
for $F \in \mathbf{F}_{\text{pruned}}$ **do**
  $\mathbf{F}_{\text{similar}} = \{F_j : R(F, F_j) > T_{\text{ransac,high}}\}_{j=1}^{N_{\text{frames}}}$
  $\mathbf{F}_{\text{nearby}} = \{F_j : |T(F) - T(F_j)| < T_{\text{time,low}}\}_{j=1}^{N_{\text{frames}}}$
  $\mathbf{F}_{\text{intra}} \coloneqq \mathbf{F}_{\text{intra}} \cup (\mathbf{F}_{\text{similar}} \cap \mathbf{F}_{\text{nearby}})$
end for

cluded from further consideration. Then, for all the remaining frames, local features are extracted and global REVV signatures are generated by aggregating these local features. The global signatures for every pair of frames are compared, resulting in a matrix of pairwise similarity scores.

Subsequently, detection of anchor frames is performed in these last three stages: (1) formation of initial anchor candidates, (2) pruning of false candidates, and (3) expansion of the candidate set to include nearby visually similar frames. The detailed description of the last three stages is presented in Algorithm 1, and the output of the algorithm is a set of anchor candidate frames $\mathbf{F}_{\text{intra}}$. The thresholds used in the algorithm are optimized separately for each news program by iterative coordinate ascent to maximize the detection accuracy over a training set. Higher values for T_{score}, $T_{\text{ransac,low}}$, and $T_{\text{ransac,high}}$ yield frame matches which are more visually similar. Higher values for N_{ransac} require each anchor frame to match a larger number of other anchor frames at different times in the video. Finally, higher values of $T_{\text{time,high}}$ require visual matches more separated in time, while higher values of $T_{\text{time,low}}$ allow for a larger temporal neighborhood in the final expansion step.

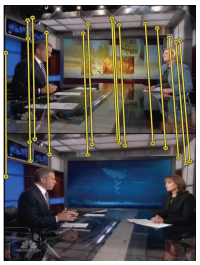


Fig. 4. Examples of matching anchor frames from different episodes of NBC Nightly News (left) and ABC World News (right). Yellow lines connect matching feature keypoints.

Algorithm 2 Inter-episode anchor detection.

$\mathbf{F}$ = frames from a video
$\mathbf{G}_{\text{anchor}}$ = anchor frames from another video
$\mathbf{F}_{\text{intra}} = \text{IntraEpisodeDetector}(\mathbf{F})$
$\mathbf{F}_{\text{other}} = \mathbf{F} \setminus \mathbf{F}_{\text{intra}}$
$\mathbf{F}_{\text{inter}} = \emptyset$
for $G \in \mathbf{G}_{\text{anchor}}$ **do**
  $\mathbf{F}_{\text{similar}} = \{F \in \mathbf{F}_{\text{other}} : S(F, G) > T_{\text{score,inter}}\}$
  $\mathbf{F}_{\text{similar,max}} = \text{TNS}(\mathbf{F}_{\text{similar}})$
  $\mathbf{F}_{\text{ransac}} = \{F \in \mathbf{F}_{\text{similar,max}} : R(F, G) > T_{\text{ransac,inter}}\}$
  $\mathbf{F}_{\text{inter}} \coloneqq \mathbf{F}_{\text{inter}} \cup \mathbf{F}_{\text{ransac}}$
end for

3.2. Intra + Inter-Episode Detection

By incorporating additional information about inter-episode visual similarity for a given news program, we can further improve detection accuracy. Challenges for the inter-episode detector are that on different days, the main anchor will wear different outfits and there may be substitute or weekend anchors. For example, Fig. 4 shows matching frames from different episodes of NBC Nightly News and ABC World News. In these examples, matching visually similar patterns in the studio background may be just as important as matching similarities in the anchor himself/herself. Certain types of anchor shots which occur once per episode, e.g., side-view shots, may be missed by the intra-episode detector but can be well detected using the inter-episode detector.

The procedure described in Algorithm 2 is carried out to detect inter-episode anchor matches $\mathbf{F}_{\text{inter}}$, starting from the intra-episode candidates $\mathbf{F}_{\text{intra}}$ from the same video and anchor frames $\mathbf{G}_{\text{anchor}}$ from another video. Then, the intra + inter-episode detector considers the union $\mathbf{F}_{\text{inter}} \cup \mathbf{F}_{\text{intra}}$ as the final result. The intra-episode detector excels at finding an-

Table 1. Comparison of preview matching performance. Types A and B refer to preview frames that occur at the start of an episode or before a commercial break, respectively.

	REVV	GIST
Memory Usage Per Video	10.3 MB	65.9 MB
Retrieval Recall: Type A	0.90	0.48
Retrieval Recall: Type B	0.93	0.62

chor frames with similarities to other anchor frames within the same episode, while the inter-episode detector effectively finds the types of anchor frames that reoccur across different episodes, e.g., side-view shots. Experimental results in Sec. 4 will show that the intra + inter-episode detector obtains the best performance.

4. EXPERIMENTAL RESULTS

4.1. Preview Matching

To evaluate the accuracy of the preview matching algorithm, we assembled a set of news videos covering 5 episodes of NBC Nightly News and 5 episodes of ABC World News from December of 2012. All matching preview and story frames in these videos were manually annotated. Keyframes are uniformly sampled at 10 frames/second from each video to ensure accurate matching even for preview shots containing large object or camera motions. Thus, 18000 keyframes are extracted for every 30-minute episode. In total, 4516 preview keyframes were detected in the 10 test videos, with 2543 Type A preview keyframes (occurring at the start of episodes) and 1973 Type B preview keyframes (occurring before commercial breaks). Each preview frame is queried against a database of story frames from the same episode. Over all the test videos, we measure the retrieval recall = (# correctly matched preview frames) / (# preview frames).

In this study, two types of global image signatures are evaluated for retrieval performance: (1) REVV signatures which are designed for efficient mobile visual search [19] and (2) GIST signatures [22, 23] which are popular in the computer vision literature and have been utilized in many image matching applications. To generate REVV signatures, we employ local SURF features [18] for its low computational cost. REVV uses a codebook of 190 centroids and a 32-bit hash for each centroid visited by the local features of an image. To generate GIST signatures, according to the recommended settings in [23], each image is resized to a 32×32 square image and a 960-dimensional GIST vector is computed. By comparing global signatures between the adaptively cropped preview region and all the video keyframes, a ranked list of story frames is generated for each preview frame. Within a shortlist of the 50 most similar story frames, geometric verification with RANSAC [20] is further applied. Empirically, a RANSAC threshold of 25 feature matches enables removal of all false positive image matches.

Table 2. Comparison of anchor detection performance.

	Recall	Precision	F-Score
REVV Intra + Inter	0.90	0.91	0.90
REVV Intra	0.87	0.90	0.88
GIST Intra	0.53	0.84	0.65

The memory usage by REVV and GIST for the 18000 keyframes per episode are listed in Table 1. Both REVV and GIST produce compact signatures for efficient memory usage. It is therefore possible to use either REVV or GIST to compare a large set of frames quickly and process many videos in parallel on any standard multi-core server that is equipped with random access memory of tens of gigabytes.

In addition to the smaller memory footprint for REVV, the main advantage of REVV over GIST for preview matching becomes evident when we examine the retrieval recall in Table 1. REVV attains significantly higher recall because it is designed to be more resilient against geometric and photometric distortions, while GIST lacks robust invariance against such distortions. To match preview and story frames accurately in news videos, it is important to use a global image signature that tolerates challenging appearance variations.

4.2. Anchor Detection

To evaluate the accuracy of the anchor detection algorithm, the 8 thresholds in Algorithm 1 and Algorithm 2 are trained on 12 episodes of NBC Nightly News (1 anchor/episode), ABC World News (1 anchor/episode), and Nightly Business Report (2 anchors/episode), and then the detection algorithm is tested on 21 episodes of those news programs. The episodes were broadcast in December of 2012 and January of 2013. Keyframes are uniformly sampled at 5 frames/second from each video. Since anchor shots have lower motion, a lower sampling rate is used in this case compared to the rate of 10 frames/second used in preview matching. The same global image signatures computed during the preview matching task can be reused for anchor detection simply by temporally subsampling the set of global image signatures by factor of 2.

Table 2 presents the recall, precision, and F-score of detecting anchor time segments. A detected segment D and ground truth segment G are paired if the normalized overlap $(D \cap G)/(D \cup G)$ exceeds 0.5. Each ground truth segment can match at most one detected segment, and vice versa. We define recall = (# correctly detected segments) / (# ground truth segments), precision = (# correctly detected segments) / (# detected segments), and F-score = $2 \times$recall$\times$precision / (recall + precision). Like in preview matching, the detection performance using REVV and GIST signatures are evaluated. From Table 2, it is clear that the REVV-based intra-episode

detector achieves good recall and precision and substantially outperforms the GIST-based intra-episode detector. When intra-episode and inter-episode detections are combined, a further improvement in performance is achieved.

5. CONCLUSIONS

We have developed efficient algorithms for matching preview and story frames and for detecting anchor appearances in a news video. Both algorithms utilize robust and memory-efficient global image signatures to recognize instances of visual similarity in the video. Experiments have shown that both algorithms are resilient against large appearance variations between visually similar frames. At the same time, the algorithms can analyze videos near real-time and require low memory footprints, so they are well suited for parallelized processing of many videos in a large archive on a multi-core server. Using these algorithms in news video content analysis, valuable sources of preview clips and important clues about story boundaries can be obtained. Future work can continue to investigate more applications of discovering visual similarity in news videos, such as removal of reoccurring commercials, identification of similar footage on different news networks, and recognition of people, places, and other interesting objects in videos. Such applications can benefit the dynamic mixing of news video clips in large archives.

6. REFERENCES

[1] Y. Kompatsiaris, B. Merialdo, and S. Lian, *TV Content Analysis*, CRC Press, Boca Ration, FL, USA, 2012.

[2] H. Zhang, Y. Gong, S. W. Smoliar, and S. Y. Tan, "Automatic parsing of news video," in *International Conference on Multimedia Computing and Systems*, May 1994, pp. 256–266.

[3] A. Hanjalic, R.L. Lagensijk, and J. Biemond, "Template-based detection of anchorperson shots in news programs," in *IEEE International Conference on Image Processing*, October 1998, pp. 148–152.

[4] Z. Liu and Q. Huang, "Adaptive anchor detection using on-line trained audio/visual model," in *SPIE Conference on Storage and Retrieval for Media Database*, January 2000, pp. 156–167.

[5] X. Gao and X. Tang, "Unsupervised video-shot segmentation and model-free anchorperson detection for news video story parsing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 9, pp. 765 – 776, September 2002.

[6] M. De Santo, P. Foggia, C. Sansone, G. Percannella, and M. Vento, "An unsupervised algorithm for anchor shot detection," in *International Conference on Pattern Recognition*, August 2006.

[7] J. Gao, M. Qi Guo, and Q.-J. Zhao, "An unsupervised anchorperson shot detection based on the distribution properties," in *International Conference on Machine Learning and Cybernetics*, August 2007, pp. 3945–3950.

[8] L. D'Anna, G. Percannella, C. Sansone, and M. Vento, "A multi-stage approach for news video segmentation based on automatic anchorperson number detection," in *International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, November 2007, pp. 229–234.

[9] C. Ma and C.-H. Lee, "Unsupervised anchor shot detection using multi-modal spectral clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2008, pp. 813–816.

[10] M. Broilo, A. Basso, and F.G.B. De Natale, "Unsupervised anchorpersons differentiation in news video," in *International Workshop on Content-Based Multimedia Indexing*, June 2011, pp. 115–120.

[11] A. Liu, S. Tang, Y. Zhang, J. Li, and Z. Yang, "A novel anchorperson detection algorithm based on spatio-temporal slice," in *International Conference on Image Analysis and Processing*, September 2007, pp. 371–375.

[12] F. Zheng, S. Li, H. Wu, and J. Feng, "Anchor shot detection with diverse style backgrounds based on spatial-temporal slice analysis," in *International Conference on Advances in Multimedia Modeling*, January 2010, pp. 676–682.

[13] M. Bertini, A. Del Bimbo, and P. Pala, "Content-based indexing and retrieval of TV news," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 503–516, April 2001.

[14] W. Xiao, W.W.Y. Ng, P.P.K. Chan, and D.S. Yeung, "L-GEM based RBFNN for news anchorperson detection with dominant color descriptor," in *International Conference on Machine Learning and Cybernetics*, July 2010, pp. 763–768.

[15] H. Lee, J. Yu, Y. Im, J.-M. Gil, and D. Park, "A unified scheme of shot boundary detection and anchor shot detection in news video story parsing," *Multimedia Tools and Applications*, vol. 51, no. 3, pp. 1127–1145, Feb. 2011.

[16] H. Chen, S. Tsai, G. Schroth, D. Chen, R. Grzeszczuk, and B. Girod, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions," in *IEEE International Conference on Image Processing*, September 2011, pp. 2609–2612.

[17] R. Smith, "An overview of the Tesseract OCR engine," in *International Conference on Document Analysis and Recognition*, September 2007, pp. 629–633.

[18] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, June 2008.

[19] D. Chen, S. Tsai, V. Chandrasekhar, G. Takacs, R. Vedantham, R. Grzeszczuk, and B. Girod, "Residual enhanced visual vector as a compact signature for mobile visual search," *Signal Processing*, vol. 93, no. 8, pp. 2316–2327, August 2013.

[20] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, June 1981.

[21] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conference on Computer Vision and Pattern Recognition*, December 2001, pp. 511–518.

[22] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[23] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid, "Evaluation of GIST descriptors for web-scale image search," in *International Conference on Image and Video Retrieval*, July 2009, pp. 1–8.