A Framework of Video Coding for Compressing Near-Duplicate Videos

Hanli Wang[1,2], Ming Ma[1,2], Yu-Gang Jiang[3], and Zhihua Wei[1,2]

[1] Department of Computer Science and Technology,
Tongji University, Shanghai, China
[2] Key Laboratory of Embedded System and Service Computing,
Ministry of Education, Tongji University, Shanghai, China
[3] School of Computer Science, Fudan University, Shanghai, China
{hanliwang,2012mingma,zhihua_wei}@tongji.edu.cn, ygj@fudan.edu.cn

Abstract. With the development of multimedia technique and social network, the amount of videos has grown rapidly, which brings about an increasingly substantial percentage of Near-Duplicate Videos (NDVs). It has been a hot research topic to retrieve NDVs for a number of applications such as copyright detection, Internet video ranking, etc. However, there exist a lot of redundancies in NDVs, and to the best of our knowledge it is an untouched research area on how to efficiently compress NDVs in a joint manner. In this work, a novel video coding framework is proposed to effectively compress NDVs by making full use of the relevance among them. Experimental results demonstrate that a significant storage saving can be achieved by the proposed NDV coding framework.

1 Introduction

The surging of the multimedia technology is exercising a great influence on our daily life. The convenience and popularization of the video capture devices give rise to a lot of creation and uploading of original videos. Naturally, the providers of network video service have to face the challenge of the video explosion. Thus, how to store and transmit videos effectively is an urgent problem that needs to solve. Especially in large-scale network video databases such as Youtube, Google Video, Yahoo Video, Youku, etc., there exist a huge amount of Near-Duplicate Videos (NDVs). NDVs can be considered as approximately identical videos that might differ in encoding parameters, photometric variations (color, lighting changes), editing operations (captions, logo insertion), or audio overlays [1]. It is a hot research topic to retrieve NDVs which is quite crucial to a number of applications such as copyright detection, video monitoring, Internet video ranking, video recommendation, etc.

At present, these NDVs are individually compressed with some video coding standard and then stored in video servers, which requires a tremendously large storage to keep them. The relevance among these NDVs has not been fully explored and utilized for a more efficient compression. If there is a method capable

C. Gurrin et al. (Eds.): MMM 2014, Part I, LNCS 8325, pp. 518–528, 2014.

of investigating the redundancy in NDVs and jointly compress the video data, much space can be saved in video servers. This is the aim of the proposed work.

Relevant to the proposed work, there are several researches dedicated to compressing similar images together. In [2], the term of set redundancy was proposed to describe the redundancy in images and the Set Redundancy Compression (SRC) algorithm was developed to jointly compress similar images. Then, a number of SRC variants were developed such as the min-max differential method [3] and the centroid-based method [4]. A comparison of SRC algorithms is made in [5]. On the other hand, Chen *et al* [6] raised a new prediction structure to divide the whole image set into different groups for compression. For grouping the similar images more efficiently, the Minimum Spanning Tree (MST) method [7] has been designed. In [8], Zou *et al* introduced the intra coding of the newly developed video coding standard High Efficiency Video Coding into image set compression and achieved an improved performance. Based on discrete cosine transform pyramid multi-level low frequency template, Li *et al* [9] proposed a method to apply subband approximation as the prediction template for similar images compression. In [10], Yue *et al* utilized the scale invariant feature transform descriptor to group and store similar images for compression.

The methods mentioned-above are designed to compress similar images, however, how to jointly compress NDVs is not exploited in the literature to the best knowledge of us. The main challenge on NDV joint compression is how to design an analysis and coding framework which can fully explore the redundancies in NDVs and then reduce these redundancies by an efficient video coding structure, which is the focus of this work. The rest of this paper is organized as follows. Related works are described in Section 2, including a brief overview of NDV retrieval and multiview video coding, which are useful to elicit the proposed NDV joint compression framework being proposed in Section 3. Experimental results are given in Section 4. Finally, Section 5 concludes the paper and presents future research prospects.

2 Related Work

A concise overview of NDV retrieval and Multiview Video Coding (MVC) is presented below. The proposed NDV joint compression framework applies NDV retrieval methodology to locate NDVs and analyze the similar images among NDVs. The MVC is extended as the prototype of the proposed NDV joint compression framework.

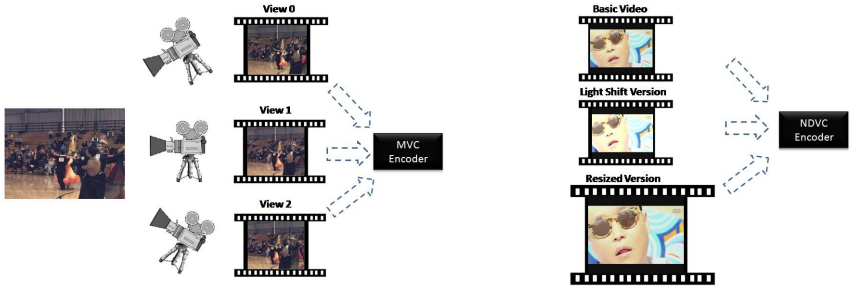
2.1 Near-Duplicate Video Retrieval

NDV retrieval has become a hot research topic in recent years, and the research efforts may be divided into two categories. One is for the retrieval speed and the other for the retrieval precision. As for faster retrieval, global features are widely adopted, such as the color [11], the edge [12], and the ordinal [13]. This category of algorithms gets a good score in retrieving videos which only have little

variations, e.g., to add subtitle or logo, or to shift the pixel lightness, etc. However, if videos are not nearly identical, the global features will be unstable and the retrieval precision is unsatisfied. Therefore, local feature-based methods are proposed for retrieval precision improvement, such as [14]-[15] to name a few. In these methods, the Bag-of-Features technique [16] is widely used. On the other hand, Sarkar *et al* [17] introduced a vector quantization-based descriptor method for NDV retrieval. A three-dimensional structure tensor model was designed for online retrieval system in [18]. In [19], Chiu *et al* proposed to speed up the NDV retrieval process by efficiently skipping unnecessary subsequence matches. Huang *et al* [20] employed a sequence of compact signatures called linear smoothing functions as the video's feature to speed up retrieval. In [21], the content fingerprint is used for NDV retrieval by making use of the spatial-temporal relevance among NDVs. Cheng *et al* [22] developed a stratification-based key frame method to enhance the precision and efficiency for copy detection.

2.2 Multiview Video Coding

Multiview videos have attracted much attention in a wide range of multimedia applications, such as three-dimensional television, free-viewpoint television, etc. A multiview video consists of video sequences of the same scenario captured by multiple cameras, but from different angles and locations, resulting in the need to store and/or transmit tremendous amounts of data. In order to compress the multiview videos effectively, Multiview Video Coding (MVC) [23] was designed for exploring not only temporal but also inter-view redundancies and thus providing higher coding performance than the independent mono-view coding. The joint compression of NDVs or in another word, Near Duplicate Video Coding (NDVC), is quite analogous to MVC if each video of NDVs is considered as one of the video views in a multiview video system. The analogy between MVC and NDVC is illustrated in Fig. 1.



(a) MVC (b) NDVC

Fig. 1. Analogy between MVC and NDVC

In order to explore both the temporal and inter-view redundancies among multiview videos, sophisticated prediction structures are employed in MVC with a typical example shown in Fig. 2. Among all views, the first view (i.e., Camera 0 in Fig. 2, also known as the base view) is coded independently; while for the other views, all frames are predicted with temporal motion estimation and/or inter-view disparity estimation.

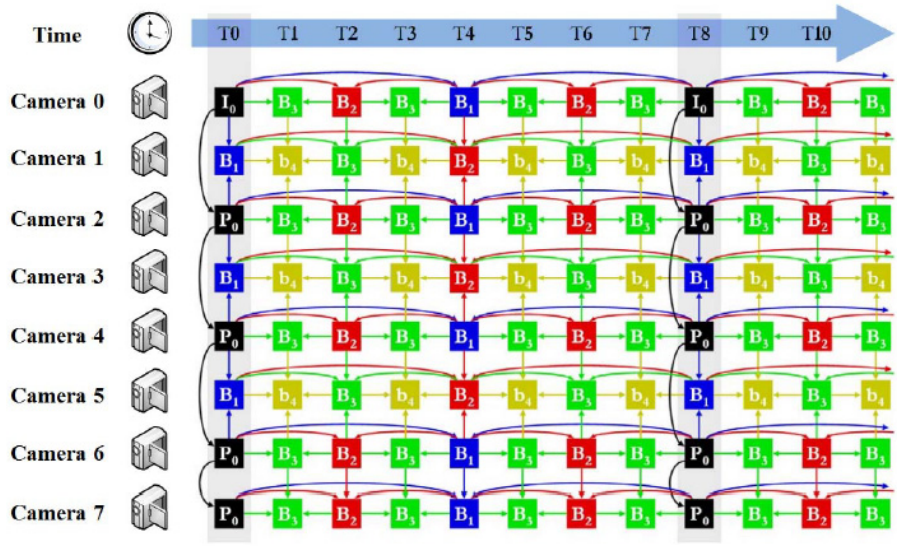


Fig. 2. A typical 8 view prediction structure in MVC [24]

3 Proposed Near-Duplicate Video Coding Framework

The proposed NDVC framework is described below, including four major components: analyzer, encoder, assembler and decoder.

3.1 Analyzer

The functionality of the analyzer is to identify NDVs and make prediction reference among them. For the convenience of discussion, the case of two NDVs is presented in the rest of this paper, which can be extended to multiple NDVs' cases. The analyzer employs some NDV detection approach to judging whether two videos are near-duplicate. If the NDVC condition holds true, the analyzer will further investigate the prediction relation between these two NDVs by generating reference indices. In other words, one video will be assigned as the basic video and the other as the dependent video which is encoded/decoded with the reference of the basic video. For each frame of the dependent video, the analyzer

will detect its most similar frame in the basic video. If the similarity score between these two frames is higher than a predefined threshold, it indicates that these two frames are similar enough to become a reference pair and the dependent frame may be coded by referencing the corresponding basic frame. However, generally speaking, the basic video and the dependent video usually exhibit an indirect reference relation. For example, the basic video and the dependent video record the same scenery by different cameras from different view points. Under such a situation, the basic video needs to be transformed with a homography matrix, and use the resultant video for reference. Therefore, the pre-processing task is also implemented by the analyzer.

3.2 Encoder

The NDVC encoder is designed by extending the encoder part of MVC as shown in Fig. 3. The main difference of NDVC encoder from that of MVC is that the currently coding frame can choose the reference frame (as signaled as the NDV reference index which is generated by the NDVC analyzer) from the corresponding basic video for prediction. Based on the basic reference frame index passed by the analyzer, the encoder performs the rate-distortion optimization-based mode decision process to choose the best coding parameters including the reference index, encoding mode, motion vectors, etc. Then, motion compensation is carried out to generate the prediction residual to remove video signal redundancies, followed by the typical Discrete Cosine Transform (DCT), Quantization (Q), entropy coding to produce the coded bitstream. The coding components of Inverse DCT (IDCT) and Inverse Q (IQ) are employed to reconstruct the video frames for subsequent reference prediction.

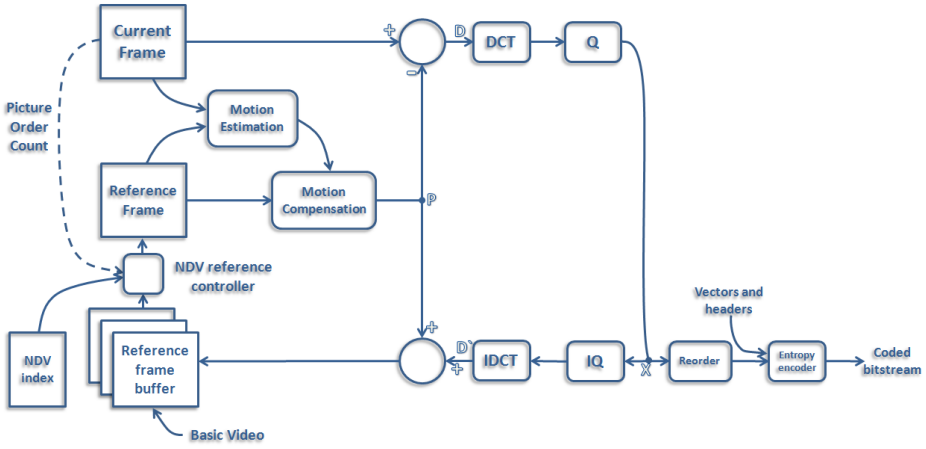
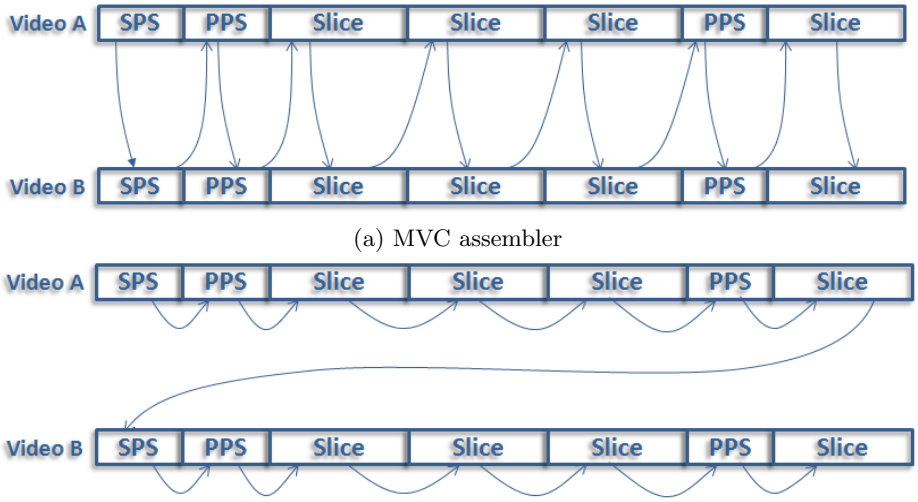


Fig. 3. Flowchart of NDVC encoder

3.3 Assembler

After encoding, each video is compressed into one bitstream. In view of the convenience of storing and decoding, the proposed NDVC framework utilizes an assembler to combine these individual bitstreams into a single bitstream as MVC does. However, the details of the NDVC assembler is different from that of MVC as shown in Fig. 4. In MVC, the inter-view reference exists at the same temporal position as show in Fig. 2. Therefore, the MVC assembling flow as illustrated in Fig. 4(a) is reasonable, since the reference frame will have been decoded ahead of the current frame to be coded/decoded. Whereas in NDVC, it is possible that every frame in the basic video can be referred by the current frame, so the assembler has to ensure that the whole basic video has been decoded before decoding the dependent video as shown in Fig. 4(b).



(a) MVC assembler

(b) NDVC assembler

Fig. 4. Assemblers of MVC and NDVC

3.4 Decoder

The NDVC decoder is shown in Fig. 5 which can be well understood by considering it as the reconstruction part of NDVC encoder as illustrated in Fig. 3.

4 Experimental Results

In order to evaluate the proposed NDVC framework, it is designed based on the MVC reference software JMVC with version of 8.5 [23]. Sixteen sets of NDVs

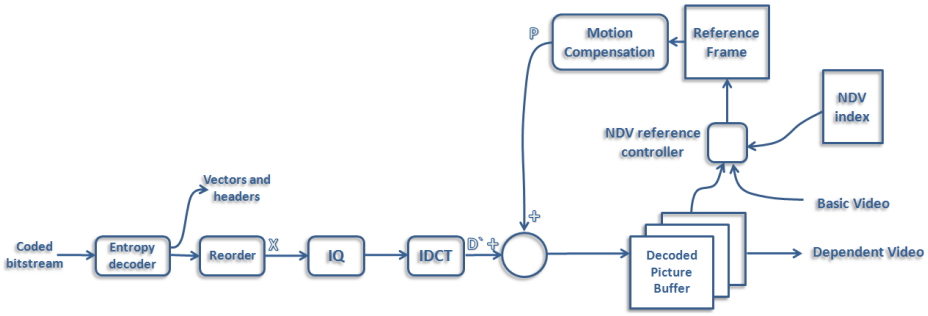
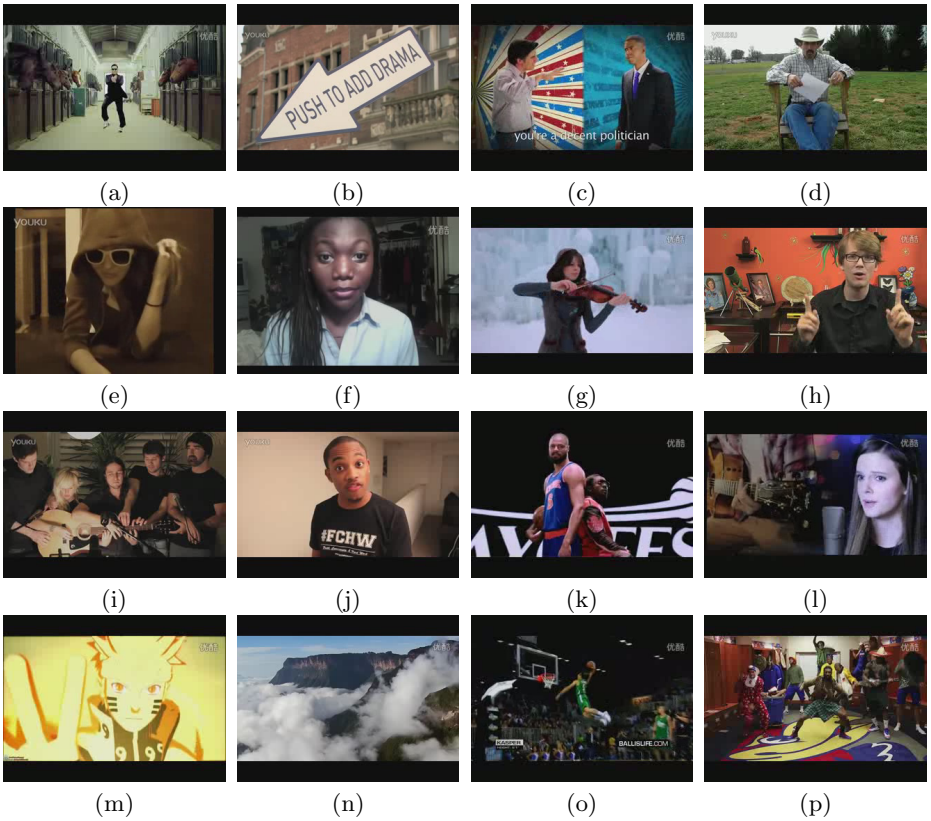


Fig. 5. Flowchart of NDVC decoder



(a) (b) (c) (d)
(e) (f) (g) (h)
(i) (j) (k) (l)
(m) (n) (o) (p)

Fig. 6. Illustration of test video sequences

are downloaded from the Youku video website [25] according to their popularity ranking by Google Zeitgeist [26] and Youku, including the Top 10 popularly queried videos at Google Zeitgeist as illustrated in Figs. 6(a)-(j) and six popular videos at Youku as shown in Figs. 6(k)-(p).

Six scenarios are applied to test the proposed NDVC framework, which are achieved by generating the dependant video with modifications on the original (i.e., basic) video, including (1) insertion of subtitle and logo, (2) addition of Gaussian white noise, (3) slowing down video play speed to half in the first 2/3 video segment and bursting up video play speed to two times in the last 1/3 video segment, (4) adjustment of 25% lightness up, (5) adjustment of 25% lightness down, and (6) resizing video resolution to half horizontally and vertically. All the test video sequences and their variations are in the format of 4:2:0 YUV. The resolution of the original videos is 320×240, and the number of frames to be coded for each video is 1575. The frame rate is 25 frames per second.

The NDV retrieval method [15] is utilized to determine whether two videos are similar enough for NDVC. Two evaluation criteria are used including Peak-Signal-to-Noise Ratio (PSNR) degradation ΔP and compression ratio $C_{\%}$, which are defined as:

$$\begin{aligned} \Delta P &= P_{NDVC} - P_{org} \\ C_{\%} &= \frac{S_{NDVC}}{S_{org}} \times 100\% \end{aligned} \tag{1}$$

where P_{NDVC} and P_{org} stand for the PSNR values of the dependant video encoded with the proposed NDVC framework and the original mono-view MVC encoder, respectively; S_{NDVC} and S_{org} are the resultant bitrates or bitstream sizes of the dependent video with the proposed NDVC framework and the original mono-view MVC encoder, respectively.

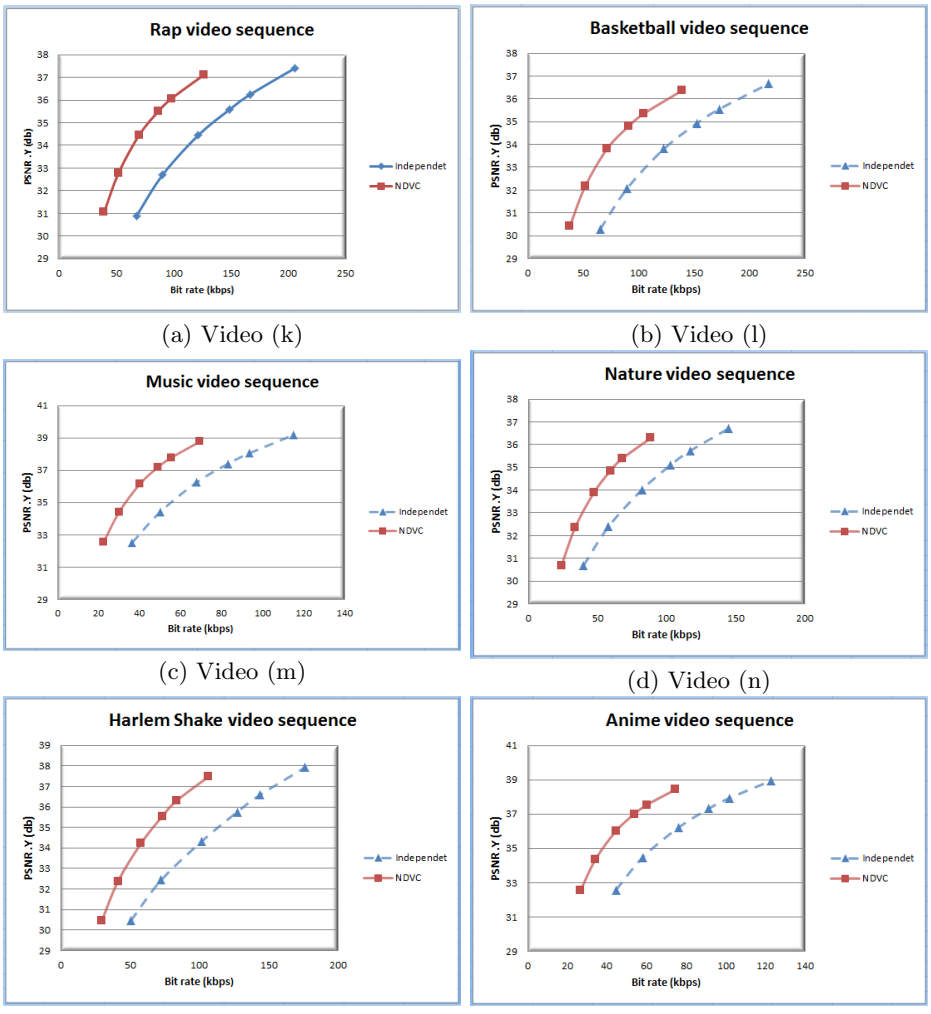
Table 1. Average results of ΔP and $C_{\%}$ under the six test scenarios

Scenario	S_{org} (kbps)	$C_{\%}$	P_{org} (dB)	ΔP (dB)
Subtitle and logo	123.36	58.82	36.89	-0.32
Gaussian noise	118.88	55.19	36.82	-0.04
Playing speed change	139.36	59.90	37.74	-0.11
Lightness up	110.16	92.51	36.38	-0.20
Lightness down	95.44	96.80	37.94	-0.02
Resize	158.64	43.24	36.84	-0.05

The summary of the experimental results are given in Table 1. Due to the space limit, only the average results on the sixteen test video sequences are presented. From the results, it can be observed that for most cases, the proposed NDVC framework will save nearly 45% of the bitrates at the cost of insignificant PSNR loss. In fact, under the current test configuration, the same Q_p value is applied to encode dependent videos by both NDVC and mono-view MVC. It has been demonstrated that by adjusting Q_p both the bitrate and PSNR

performances resulted from the proposed NDVC can be better than that of mono-view MVC, which is to be presented below. On the other hand, when dealing with the test scenario of lightness shifting, the results are not so good since pre-processing lightness shifting is currently not handled by the NDVC analyzer, which is open as one of our future research directions.

In order to further illustrate the advantage of the proposed NDVC framework and considering the space limit, the comparison of Rate-Distortion (R-D) performances of six video sequences including Videos (k)-(p) under the 'Subtitle and logo' scenario is shown in Fig. 7, where the R-D curves of NDVC and the independent mono-view MVC are both plotted for comparison. From the results,



(a) Video (k) (b) Video (l)

(c) Video (m) (d) Video (n)

(e) Video (o) (f) Video (p)

Fig. 7. Comparison of R-D performances between mono-view MVC and NDVC

it be easily observed that the proposed NDVC framework achieves significantly better R-D performances than independent video coding. Similar observations can be made for other test video sequences and scenarios.

5 Conclusion and Future Work

In this paper, a novel video coding framework NDVC is proposed to jointly compress NDVs aiming to effectively trimming down the storage of video servers, including four major parts: analyzer, encoder, assembler and decoder. Sixteen popular video sets have been applied to evaluate the proposed NDVC framework. The experimental results have demonstrated that the proposed NDVC framework is very efficient in reducing the redundancy in NDVs and thus saving the storage or network bandwidth resources.

In the future, the NDVC analyzer will be further enhanced by incorporating more pre-processing functionalities, such as adjustment of lightness and calculation of homography transform matrix, for generating better references. It is also desired to design more efficient approaches to locating similar frames between basic and dependent videos. Within the proposed NDVC framework, fast NDVC coding technologies and rate control algorithms are also preferred to improving the NDVC performances.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 61102059, the "Shu Guang" project of Shanghai Municipal Education Commission and Shanghai Education Development Foundation under Grant 12SG23, the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, the Program for New Century Excellent Talents in University of China under Grant NCET-10-0634, the Fundamental Research Funds for the Central Universities under Grants 0800219158 and 1700219104, and the National Basic Research Program (973 Program) of China under Grant 2010CB328101.

References

1. Cherubini, M., Oliveira, R., de Oliveira, N.: Understanding Near-Duplicate Videos: A User-Centric Approach. In: ACM MM 2009, pp. 35–44 (2009)
2. Karadimitriou, K.: Set Redundancy, the Enhanced Compression Model, and Methods for Compressing Sets of Similar Images. PhD Thesis. Louisiana State University (1996)
3. Karadimitriou, K., Tyler, J.M.: Min-Max Compression Methods for Medical Image Databases. In: ACM SIGMOD 1997, pp. 47–52 (1997)
4. Karadimitriou, K., Tyler, J.M.: The Centroid Method for Compressing Sets of Similar Images. Pattern Recognition Lett. 19(7), 585–593 (1998)
5. Samy, A.-A., Abdelhalim, G.: A Comparison of Set Redundancy Compression Techniques. EURASIP Journal Applied Signal Process., 1–13 (2006)
6. Chen, C.-T., Chen, C.-S., Chung, K.-L., Lu, H., Tang, G.Y.: Image Set Compression through Minimal Cost Prediction Structure. In: IEEE ICIP 2004, pp. 1290–1292 (2004)

7. Nielsen, C., Li, X.: MST for Lossy Compression of Image Sets. In: DCC 2006, p. 463 (2006)
8. Zou, R., Au, O.C., Zhou, G., Li, S., Sun, L.: Image Set Modeling by Exploiting Temporal-Spatial Correlations and Photo Album Compression. In: APSIP ASC 2012, pp. 1–4 (2012)
9. Li, S., Au, O.C., Zou, R., Sun, L., Dai, W.: Similar Images Compression based on DCT Pyramid Multi-level Low Frequency Template. In: MMSP 2012, pp. 255–259 (2012)
10. Yue, H., Sun, X., Wu, F., Yang, J.: SIFT-based Image Compression. In: IEEE ICME 2012, pp. 473–478 (2012)
11. Kasutani, E., Yamda, A.: The MPEG-7 Color Layout Descriptor: A Compact Image Feature Description for High-Speed Image/Video Segment Retrieval. In: IEEE ICIP 2001, pp. 674–677 (2001)
12. Bertini, M., Del Bimbo, A., Nunziati, W.: Video Clip Matching Using MPEG-7 Descriptors and Edit Distance. In: Sundaram, H., Naphade, M., Smith, J.R., Rui, Y. (eds.) CIVR 2006. LNCS, vol. 4071, pp. 133–142. Springer, Heidelberg (2006)
13. Kim, C., Vasudev, B.: Spatiotemporal Sequence Matching for Efficient Video Copy. IEEE Trans. Circuits Syst. Video Technol. 15(1), 127–132 (2005)
14. To, J.L., Chen, L., Joly, A., Laptev, I., Buisson, O., Brunet, V.G., Boujemaa, N., Stentiford, F.: Video Copy Detection: A Comparative Study. In: ACM CIVR 2007, pp. 371–378 (2007)
15. Zhao, W., Wu, X., Ngo, C.-W.: On the Annotation of Web Videos by Efficient Near-Duplicate Search. IEEE Trans. Multimedia 12(5), 448–461 (2010)
16. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual Categorization with Bags of Keypoints. In: ECCV SLCV 2004, pp. 1–22 (2004)
17. Sarkar, A., Singh, V., Ghosh, P., Manjunath, B.S., Singh, A.: Efficient and Robust Detection of Duplicate Videos in A Large Database. IEEE Trans. Circuits Syst. Video Technol. 20(6), 870–885 (2010)
18. Zhou, X., Chen, L.: ASVTDECTOR: A Practical Near Duplicate Video Retrieval System. In: ICDE 2013, pp. 1348–1351 (2013)
19. Chiu, C.-Y., Li, S.-Y., Hsieh, C.-Y.: Video Query Reformulation for Near-Duplicate Detection. IEEE Trans. Inf. Forensics Security 14(4), 1220–1233 (2012)
20. Huang, Z., Shen, H.T., Shao, J., Cui, B., Zhou, X.: Practical Online Near-Duplicate Subsequence Detection for Continuous Video Streams. IEEE Trans. Multimedia 12(5), 386–398 (2010)
21. Esmaeili, M.M., Fatourechi, M., Ward, R.K.: A Robust and Fast Video Copy Detection System Using Content-Based Fingerprinting. IEEE Trans. Inf. Forensics Security 6(1), 213–226 (2011)
22. Cheng, X., Chia, L.-T.: Stratification-Based Keyframe Cliques for Effective and Efficient Video Representation. IEEE Trans. Multimedia 13(6), 1333–1342 (2011)
23. Vetro, A., Pandit, P., Kimata, H., Smolic, A., Wang, Y.-K.: Joint Draft 8.0 on Multiview Video Coding. JVT-AB204 (2008)
24. MerKle, P., Müller, K., Smolic, A., Wiegand, T.: Efficient Compression of Multi-View Video Exploiting Inter-View Dependencies based on H.264/MPEG4-AVC. In: ICME 2006, pp. 1717–1720 (2006)
25. Youku Webiste, `http://www.youku.com/`
26. Google Zeitgeist, `http://www.google.com/zeitgeist/`