NON-IDENTICAL DUPLICATE VIDEO DETECTION USING THE SIFT METHOD

Karthikeyan Vaiapury[1], Pradeep K. Atrey[1], Mohan S. Kankanhalli[1] and Kalpathi Ramakrishnan[2]

[1] School of Computing, National University of Singapore, Republic of Singapore
[2] Department of Electrical Engineering, Indian Institute of Science, Bangalore

Keywords: NID video detection, SIFT.

Abstract

Non-Identical Duplicate video detection is a challenging research problem. Non-Identical Duplicate video are a pair of videos that are not exactly identical but are almost similar. In this paper, we evaluate two methods - Keyframe-based and Tomography-based methods to determine the Non-Identical Duplicate videos. These two methods make use of the existing scale based shift invariant (SIFT) method to find the match between the key frames in first method, and the cross-sections through the temporal axis of the videos in second method. We provide extensive experimental results and the analysis of accuracy and efficiency of the above two methods on a data set of Non-Identical Duplicate video-pair.

1 Introduction

Duplicate detection in multimedia content is a subject of active research within the community. The duplicate media content can exist because of two reasons - first, a copy of a video for transcoding purposes or for illegal copying of potential content; second, the consumers more often shoot multiple photos and videos of the same scene. The problem of duplicate detection in the first case is the problem of matching exactly two similar media contents, the solutions for which have been proposed using various digital signature/watermarking based methods [4], [2]. In the second case, the duplicate detection is performed by matching two media contents which are not exactly identical but almost similar (such media content are called “Non-Identical Duplicate” [5]).

The detection of Non-Identical Duplicate media content is useful for retrieval purposes such as QBE (Query by Example). For example, one may be interested in finding all the related news videos that are Non Identical Duplicate (NID) which has Abdul Kalam as the focus.

In this paper, we address the problem of detecting Non-Identical Duplicate videos. The video is a sequence of frames that have a high degree of temporal correlation among them. Each frame is an image in the two-dimensional spatial plane. The extra time dimension in video with several additional properties makes the detection of Non-Identical Duplicates in video different from that in images. We evaluate two methods for finding Non-Identical Duplicate videos - Keyframe-based method and Tomography-based method, both use the Lowe’s SIFT method [8] which is highly effective in identifying and matching the interest points even for transformed images. We have performed extensive experiments and have provided the analysis of accuracy and efficiency of such approach in detecting the Non-Identical Duplicate videos.

The paper is organized as follows. We provide the related work in section 2. In section 3, we present the proposed method. Section 4 presents the extensive experimental results and analysis. Finally, we conclude the paper with a discussion on future work in section 5.

2 Related work

The problem of detection of Non-Identical Duplicate media content has been studied in the past mostly in context of images [5, 3, 6, 7]. The general methods of matching two video clips include the signature/watermarking based [5], [6], [2] and the distance based [4] matching methods. However, the accuracy of such method highly depends on the deployed features and moreover, they are not suitable for matching non-identical duplicate videos due to their sensitivity to the change in color, brightness, frame format and the transformation such as scaling, rotation, down sampling etc. Significant amount of works have been done on searching near-replica, non-identical duplicates for images [8, 3, 7]. [7] used PCA SIFT based on SIFT [8] which is resistant to transformations such as scaling, rotation, down sampling etc. However the problem of finding near replica using scale invariant method has been addressed only for the images.

In this work, we have addressed the Non-Identical Duplicate problem for videos using the keyframe-based and the tomography-based SIFT methods. The SIFT method [8] is used for finding similarity match between any two frames and it provides significantly accurate results, though it is computational expensive. This method is used because of its robustness to variation in scale, rotation, affine distortion and 3D viewpoint etc. The steps used in generating the features of video key frames include a) Finding the scale and orientation invariant interest points using DoG (Difference of Gaussian function), b) Selecting the key points based on stability, c) Assigning the orientation, and d) Key point descriptor.

3 Proposed work

We evaluate two methods - Keyframe-based and Tomography-based methods to find whether or not the given videos are Non-Identical Duplicates. Both the methods use SIFT-method for matching two (key-frames) in Keyframe-based method and (cross-sections) in Tomography-based method.

3.1 Keyframe-based method

In order to determine whether or not the given two videos are Non-Identical Duplicate, the Keyframe-based method works as follows -

1. Given the two input videos (say V1 and V2), we extract the key frames from both the videos using the simple differencing and thresholding method similar to what is used in [2]. Let KF1 = $\{k_{1i}\}, 1 \leq i \leq n_1$ and KF2 = $\{k_{2j}\}, 1 \leq j \leq n_2$ be the sets of key frames of videos V1 and V2, respectively; where n_1 and n_2 are the total number of key frames in videos V1 and V2, respectively.

2. The similarity matches between all the pairs (k_{1i}, k_{2j}), $1 \leq i \leq n_1, 1 \leq j \leq n_2$ are obtained using the SIFT method [8]. As a result of this, we obtain a matrix M_{ij}, $1 \leq i \leq n_1, 1 \leq j \leq n_2$. The match score M_{ij} is computed using the following equation -

$$M_{ij} = 2 \times \frac{m_{ij}}{P_i + P_j} \tag{1}$$

   where, m_{ij} is the number of match points between the frames i and j, and P_i and P_j are the number of key points found using SIFT method for the frame i and frame j, respectively.

3. Finally, we obtain the Overall Match Score (OMS) between the two videos by averaging all the match scores M_{ij} in a diagonal band of a given width. Precisely, the OMS_k for a band of width k is computed as -

$$OMS_k = mean(M_{ij}) \tag{2}$$

   where, $1 \leq i \leq n_1, 1 \leq j \leq n_2$ and $i = j - (k-1)/2$.

In case, $n_1 \neq n_2$, we repeat the diagonal band until the last element of n_1 or n_2 whichever is higher, and compute the Overall Match Score by averaging the match scores of all the bands.

3.2 Tomography-based method

We also explored the use of tomography-based [1] matching method to find the Non-Identical Duplicate videos.

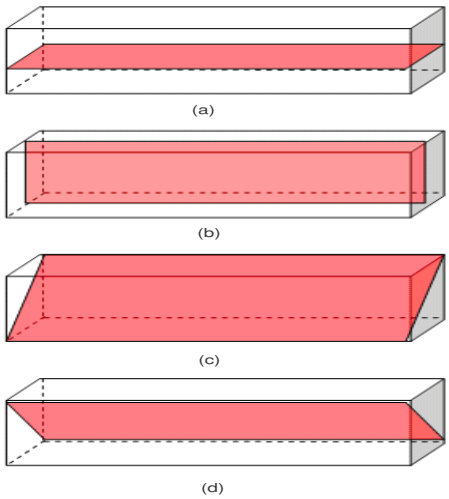

Figure 1: Four cross-sections used in our experiments (a) Middle horizontal (b) Middle vertical (c) Diagonal 1 and (d) Diagonal 2

Tomography refers to finding the cross-sections through the temporal axis of the video. In this method, instead of using key frames, we used the four (middle-horizontal, middle-vertical, and the two diagonal) cross-sections through the temporal axis of the two videos as shown in figure 1, and then compared them with the similar cross-sections of the other video using SIFT-method. The Overall Match Score (OMS) is computed by averaging the individual match scores of all the four cross-sections matching.

4 Experimental Results

The extensive experiments are performed to analyze the accuracy and efficiency of SIFT method in detecting the Non-Identical Duplicate videos. For our experiments, we used Pentium-IV 2.4 GHz with 512 MB RAM. We prepared a data-set of 16 Non-Identical Duplicate video-pair (shown in figure 2). All the videos were shot using a Fujifilm digital camera (Model A210) with 320×240 resolution. Note that since no standard data set on NID-video pair is publicly available, we have used our own data set. The past works have used the "Run-Lola-Run" data set which provides only NID-images not the videos. We also report results on the "Run-Lola-Run" data set considering the NID-images as a representative to the video.

4.1 Experiments on our data set

In keyframe-based SIFT method, we first extract the key frames from each video. The total number of frames in each video and the number of key frames extracted from them are shown

Figure 2: The key frames of 16 Non-Identical Duplicate video pair of our data set

Table 1: Details of Non-Identical Duplicate videos used in our experiment

NID Video pair	Total number of frames	Duration of clips (in seconds)	Number of key frames (Set 1) n_1, n_2	Number of key frames (Set 2) n_1, n_2
PgpHostel	99, 89	9, 8	23, 16	6,4
BusNuh	40, 40	3, 3	10, 10	1,2
BusRoad	50, 70	4, 6	14, 22	5,8
CampusMap	120, 100	12, 9	12, 5	2,1
CarPark	70, 90	6, 8	13, 16	5,6
Corridor	60,70	5, 6	3, 3	1,1
KartikCanteen	60, 60	5, 5	13, 15	4,5
SocLift	30, 30	3, 3	3, 1	1,1
OurLab	80, 100	7, 9	21,14	12, 8
PgpRoad	70, 30	6, 3	10, 17	3, 7
PradeepCanteen	120, 130	14, 16	43, 69	20, 30
PlayGround	90, 80	8, 7	5, 3	3,1
FoodCanteen	110, 130	10, 12	30, 37	10, 11
SocS16	100, 90	9, 8	24, 28	11, 12
TeaCanteen	60, 40	5, 4	7, 4	2, 1
Laptop	40, 60	4 ,5	3, 4	1,2

in Table 1. We have obtained two sets (Set 1 and Set 2) of key frames using different threshold values. This is done with the aim of analyzing how much computational efforts can be saved while maintaining the accuracy. We compared each video-pair using SIFT method and computed the Overall Match Score (OMS) using equation (2). The value of width k of band is chosen as 1 and 3. In other words, $k = 1$ allows to compare each keyframe in one video with the corresponding key frame in the other video; while with $k = 3$, we compare each key frame in one video with the one previous and one next keyframe in addition to the corresponding key frame of the other video.

We used a Bayesian classifier to categorize a given input video-pair into one of the two classes - NID pair and Non-NID pair. The Bayesian classifier is trained based on Overall Match Score (OMS) obtained for a set of NID-pair and Non-NID pair.

We have performed experiment using the Keyframe-based SIFT method on all the video pairs (shown in figure 2) of original resolution i.e. 320×240. The same experiment is performed on the video pairs by down-grading them to lower resolutions i.e. 160×120, 80×60, 40×30 and 20×15. We did this experiment to observe how the accuracy and efficiency are affected by matching low resolution video pairs. The Overall Match Score is computed for the two different values of width k of the diagonal band (refer to section 3.1), i.e. for $k = 1$ and $k = 3$. The results for OMS_1 (for $k = 1$) and for OMS_3 (for $k = 3$) are reported in Table 2 and Table 3, respectively. In Table 2 and Table 3, each entry for a particular resolution of the video pair has three values - Overall Match Score (OMS_k) (multiplied by 10^4), the time (T_k) taken in computing this score and the probability (P_k) of this pair belonging to a NID-pair class, with k being the width of diagonal band.

We also performed experiment on the video pairs of original resolution (i.e. 320×240) by using the Set 2 of key frames (as described in Table 1). The results for this experiment are reported in Table 4.

The experiments are also performed for Tomography method (as described in section 3.2). The results are shown in Table 5.

The overall observations from the obtained results are-

1. From our experiments, we clearly found that SIFT based method is effective and works well for detecting Non-Identical Duplicate videos.

Table 2: Overall Match Score (OMS1 with $k = 1$) for NID video pairs (with Set 1 of key frames) at varying resolutions using Keyframe-based SIFT method

NID Video pair	320×240			160×120			80×60			40×30			20×15		
	OMS_1	T_1 (Sec)	P_1	OMS_1	T_1 (Sec)	P_1	OMS_1	T_1 (Sec)	P_1	OMS_1	T_1 (Sec)	P_1	OMS_1	T_1 (Sec)	P_1
PgpHostel	508	158	0.9998	823	136	0.9999	743	122	1.0000	693	109	0.9999	517	33	0.9995
BusNuh	55	42	0.9423	98	33	0.9970	109	27	0.9847	358	25	0.9993	0	2	0.0014
BusRoad	31	248	0.7625	81	80	0.9637	66	72	0.9482	143	66	0.9929	0	1	0.0014
CampusMap	227	50	0.9980	220	43	0.9966	131	18	0.9902	0	12	0.0026	0	2	0.0014
CarPark	408	108	0.9996	789	55	0.9999	964	44	1.0000	842	25	0.9999	220	23	0.9971
Corridor	3744	4	1.0000	4822	4	1.0000	6631	2	1.0000	6274	2	1.0000	4111	2	1.0000
KartikCanteen	706	42	0.9999	800	20	0.9999	478	16	0.9996	296	11	0.9988	0	1	0.0014
SocLift	6266	3	1.0000	6950	3	1.0000	1268	3	1.0000	1182	3	1.0000	222	3	0.9971
OurLab	369	139	0.9994	125	36	0.9858	0	16	0.0027	0	10	0.0026	0	1	0.0014
PgpRoad	559	168	0.9998	491	45	0.9996	382	36	0.9993	167	30	0.9951	0	1	0.0014
PradeepCanteen	61	2536	0.9551	92	2408	0.9733	35	727	0.7409	0	614	0.0026	0	2	0.0014
PlayGround	3978	12	1.0000	5140	10	1.0000	5886	9	1.0000	6477	6	1.0000	0	1	0.0014
FoodCanteen	2503	366	1.0000	1383	343	1.0000	1831	326	1.0000	777	221	1.0000	0	2	0.0014
SocS16	3235	158	1.0000	2981	152	1.0000	2978	148	1.0000	3007	141	1.0000	0	1	0.0014
TeaCanteen	890	17	1.0000	584	14	0.9997	648	10	0.9999	1052	9	1.0000	500	6	0.9995
Laptop	2763	6	1.0000	4668	5	1.0000	4283	4	1.0000	5576	3	1.0000	746	2	1.0000

2. Keyframe-based method performs with a decent accuracy and is much faster than the simple approach of comparing two videos by matching each frame of one video with the each frame of the other video within a certain width of band. We verified it by comparing the Non-Identical Duplicate pair of the video 'PgpHostel'. The computation time for matching (89 frames of PgpHostel1 and 99 frames of PgpHostel2) is found around 270 minutes, which is significantly greater than the average computation time recorded for keyframe-based method.

3. We have observed (from Table 2 and Table 3) that the computation time significantly decreases for the lower resolution video pairs (as also can be seen in figure 3), and it also provides reasonably accurate results upto a certain resolution. This is because, by reducing the resolution of the frame, the number of key points that are invariant are also reduced but lesser number of match points are reduced. As can be seen from figure 4, for OMS_1, the accuracy remains more than 83% upto resolution 40×30 and it degrades to 37% with resolution 20×15. This indicates that SIFT-based method works even for matching the low-resolution video pairs.

4. It is also observed that the overall accuracy (for our data set) is found to be better with $k = 3$ compared to that with $k = 1$. As can be seen in figure 4, for video pairs of resolution 40×30, the accuracy for OMS_3 is 93.75% which is higher compared to the accuracy for OMS_1 which is 81.25%.

5. It is also interesting to note that, for Set 2 of key frames, the SIFT method performed well i.e. with accuracies of 87.5% (for OMS_1) and 100% (for OMS_3) (Refer to Table 4). Also, it provides a significant amount of gain in the computation time (i.e. around 10 times faster).

6. After comparing the Tomography-based method with the Keyframe-based method, we found that Tomography-based method does not perform well for the data set

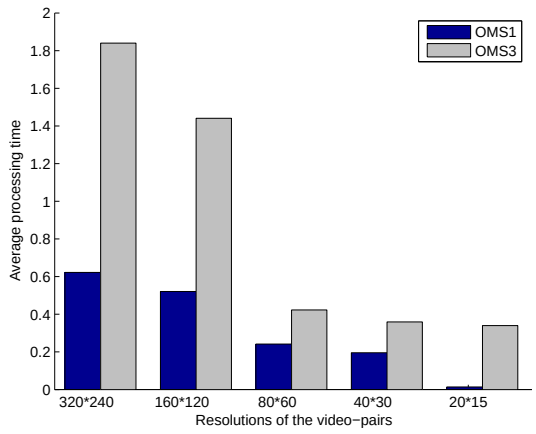

Figure 3: Average processing time (in seconds) for matching a single frame-pair of the videos of different resolutions

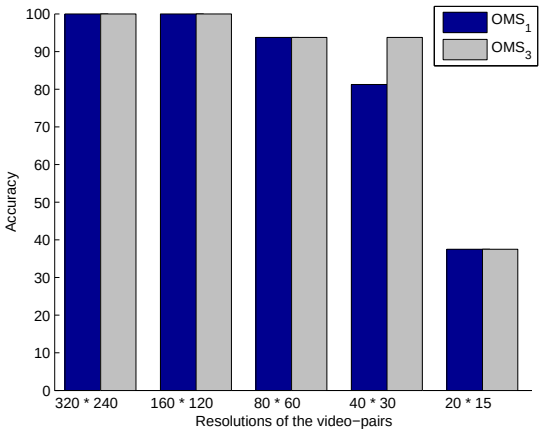

Figure 4: Accuracy of NID video pair detection vs. Resolution of video pairs

Table 3: Overall Match Score (OMS3 with $k = 3$) for NID video pairs (with Set 1 of key frames) at varying resolutions using Keyframe-based SIFT method

NID Video pair	320×240			160×120			80×60			40×30			20×15		
	OMS_1	T_1 (Sec)	P_1	OMS_1	T_1 (Sec)	P_1	OMS_1	T_1 (Sec)	P_1	OMS_1	T_1 (Sec)	P_1	OMS_1	T_1 (Sec)	P_1
PgpHostel	485	470	0.9993	787	402	0.9996	826	263	1.0000	611	243	0.9998	458	107	0.9998
BusNuh	63	110	0.8848	125	80	0.9858	180	72	0.9949	406	67	0.9994	0	2	0.0014
BusRoad	36	740	0.5937	66	234	0.9337	68	102	0.9472	221	99	0.9971	0	1	0.0014
CampusMap	252	148	0.9960	278	105	0.9980	171	27	0.9942	94	24	0.9776	0	2	0.0014
CarPark	502	320	0.9994	900	110	1.0000	916	65	1.0000	960	35	1.0000	658	28	1.0000
Corridor	3490	11	1.0000	4581	6	1.0000	5804	3	1.0000	6142	2	1.0000	692	2	1.0000
KartikCanteen	1179	124	1.0000	1403	55	1.0000	920	36	1.0000	833	23	0.9999	0	1	0.0014
SocLift	786	9	0.9999	1647	3	1.0000	1612	3	1.0000	1479	3	1.0000	1479	3	1.0000
OurLab	367	395	0.9985	131	112	0.9902	0	36	0.0027	0	29	0.0026	0	1	0.0014
PgpRoad	718	502	0.9998	669	127	0.9998	371	56	0.9992	240	47	0.9976	0	1	0.0014
PradeepCanteen	512	7520	0.9994	77	6875	0.9545	43	1263	0.8335	60	1120	0.9327	0	2	0.001
PlayGround	3888	24	1.0000	5009	22	1.0000	5696	19	1.0000	5909	12	1.0000	0	1	0.0014
FoodCanteen	2490	1082	1.0000	1359	772	1.0000	1766	451	1.0000	717	359	0.9999	0	2	0.0014
SocS16	3852	470	1.0000	3075	433	1.0000	3032	306	1.0000	2745	230	1.0000	0	1	0.0014
TeaCanteen	402	50	0.9988	507	48	0.9996	473	45	0.9996	1381	41	1.0000	1031	41	1.0000
Laptop	3920	17	1.0000	4905	11	1.0000	5686	9	1.0000	6153	6	1.0000	769	6	1.0000

Table 4: Overall Match Score (OMS) for different pair of videos (for Set 2 of key frames) at resolution 320×240 using Keyframe-based SIFT method

NID Video pair	OMS_1	T_1 (Sec)	P_1	OMS_3	T_3 (Sec)	P_3
PgpHostel	467	11	0.9997	484	33	0.9997
BusNuh	23	2	0.1764	63	6	0.9201
BusRoad	33	21	0.4973	36	63	0.6858
CampusMap	160	2	0.9887	348	6	0.9991
CarPark	437	7	0.9996	324	21	0.9989
Corridor	486	2	0.9998	486	2	0.9997
KartikCanteen	726	6	1.0000	714	17	1.0000
SocLift	2059	2	1.0000	2059	2	1.0000
OurLab	401	45	0.9994	390	135	0.9994
PgpRoad	847	16	1.0000	592	48	0.9999
PradeepCanteen	56	280	0.8382	467	835	0.9997
PlayGround	1137	3	1.0000	3888	9	1.0000
FoodCanteen	1323	14	1.0000	1388	40	1.0000
SocS16	1622	15	1.0000	1686	45	1.0000
TeaCanteen	684	2	1.0000	356	6	0.9992
Laptop	1411	2	1.0000	2079	6	1.0000

OMS_1: Accuracy = 87.5%, OMS_3: Accuracy = 100%

Table 5: Overall Match Score (OMS) for different pair of videos (for Set 1 of key frames) at resolution 320×240 using Tomography-based SIFT method

NID Video pair	OMS_1	T_1 (Sec)
PgpHostel	15	40
BusNuh	10	30
BusRoad	9	32
CampusMap	6	45
CarPark	14	36
KartikCanteen	8	34
SocLift	5	32
OurLab	12	40
PgpRoad	9	34
PradeepCanteen	10	47
PlayGround	11	36
FoodCanteen	5	32
SocS16	8	30
TeaCanteen	7	30
Laptop	15	32

provided in figure 2. Using Tomography-based method, the average OMS for NID video-pair and Non-NID video-pair was found 9.6 and 5.8, respectively, which was not good enough to distinguish between NID and Non-NID video-pair (Refer to Table 5). This might be because these videos are shot by keeping the camera in hand, which constantly changes the view point. In such videos, finding the correctly matching cross-section in two videos is difficult.

7. We also computed the Overall Match Score for Non-NID videos using SIFT-based method. The results are provided in Table 6. The results clearly show that SIFT-based method can detect Non-NID video pair with a significantly high accuracy (82% and 91%, with bandwidth k = 1 and k = 3, respectively).

Table 6: Overall Match Score (OMS) for Non-NID pair of videos (for Set 1 of key frames) at resolution 320×240 using Keyframe-based SIFT method

Video 1 (n_1)	Video 2 (n_1)	OMS_1 (Sec)	T_1	P_1	OMS_3 (Sec)	T_3	P_3
BusNuh (10)	BusRoad (14)	43	70	0.8943	69	201	0.9072
BusRoad (14)	CampusMap (12)	8	82	0.0839	8	240	0.1083
CampusMap (12)	CarPark (13)	8	68	0.0839	8	194	0.1083
CarPark (13)	Corridor (3)	9	19	0.0433	10	55	0.0542
Corridor (3)	KartikCanteen (13)	0	19	0.0039	3	55	0.2778
OurLab (21)	SoCLift (3)	14	20	0.0536	11	58	0.0330
SoCLift (3)	PlayGround (5)	0	7	0.0039	5	7	0.2076
PlayGround (5)	PradeepCanteen (43)	4	110	0.3039	4	305	0.2427
PradeepCanteen (43)	FoodCanteen (30)	8	612	0.0839	9	1804	0.0796
FoodCanteen (30)	SoCS16 (24)	5	340	0.2461	3	1270	0.2778
SoCS16 (24)	TeaCanteen (7)	26	84	0.6393	28	245	0.3646

n_1 indicates the number of key frames in the video
OMS_1: Accuracy = 82%, OMS_3: Accuracy = 91%

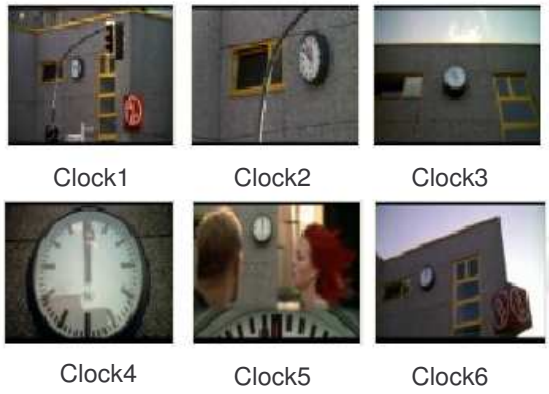

Figure 5: Run-Lola-Run data set

Table 7: Overall Match Score (OMS) for "Run-Lola-Run" data set using Keyframe-based SIFT method

Representative Image 1	Representative Image 2	OMS_1	P_1
Clock1	Clock2	10000	1.0000
Clock1	Clock3	170	0.9960
Clock1	Clock4	13	0.0213
Clock1	Clock5	0	0.0029
Clock1	Clock6	125	0.9918
Clock2	Clock3	170	0.9960
Clock2	Clock4	13	0.0213
Clock2	Clock5	0	0.0029
Clock2	Clock6	125	0.9918
Clock3	Clock4	0	0.0029
Clock3	Clock5	80	0.9766
Clock3	Clock6	97	0.9852
Clock4	Clock5	117	0.9904
Clock4	Clock6	46	0.9105
Clock5	Clock6	137	0.9934

OMS_1: Accuracy = 66%

4.2 Experiments on "Run-Lola-Run" data set

We have also performed experiment to test the SIFT based method on the "Run-Lola-Run" dataset [9]. The representative images (of resolution 459×366) of this data set are shown in figure 5. We obtained an accuracy of 66% in the NID-image pairs (Refer to Table 7). We observed that the Clock4 image is falsely detected as Non-NID image with all images except the Clock5 and Clock6.

5 Conclusions

In this paper, we have evaluated two methods - Keyframe-based method and Tomography-based method that uses a SIFT method, for the detection of Non-Identical Duplicate videos. Keyframe-based SIFT method provide significantly accurate results in the reasonable amount of computation time, while Tomography-based SIFT method does not perform well in the case of videos which has shaking artifacts. The future work will be to further investigate how Keyframe-based method and Tomography-based can be used to achieve better accuracy and efficiency on large size videos of different scales and of significantly different durations. It would also be interesting to explore the use of other methods such as wavelet-based method for detecting Non-Identical Duplicate video pair.

References

[1] A. Akutsu and Y. Tonomura. Video tomography : An efficient method for camera work extraction and motion analysis. In *ACM International Conference on Multimedia*, 1994.

[2] Pradeep K. Atrey, Yan Wei-Qi, and Mohan Kankanhalli. A scalable signature scheme for video authentication. In *Journal of Multimedia Tools and Applications*, 2006. To appear.

[3] Edward Chang, Chen Li, James Wang, Peter Mork, and Gio Wiederhold. Searching near-replicas of images via clustering. In *SPIE Multimedia Storage and Archiving Systems IV*, volume 3846, pages 281–292, 1999.

[4] Arun Hampapur and Ruud M. Bolle. Comparison of distance measures for video copy detection. In *IEEE International Conference on Multimedia and Expo*, 2001.

[5] Alejandro Jaimes, Shih-Fu Chang, and Alexander C. Loui. Duplicate detection in consumer photography and news video. In *ACM International Conference on Multimedia*, 2002.

[6] Alejandro Jaimes, Shih-Fu Chang, and Alexander C. Loui. Detection of non-identical duplicate consumer photographs. In *Pacific Rim International Conference on Multimedia*, volume 1, pages 16–20, 2003.

[7] Yan Ke, R. Suthankar, and L. Huston. Efficient near-duplicate detection and sub-image retrieval. In *ACM International Conference on Multimedia*, 2004.

[8] David G. Lowe. Distinctive image features from scale invariant key points. In *International Journal of Computer Vision*, 2004.

[9] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision*, volume 2, pages 1470–1477, October 2003.