ROBUST VIDEO FINGERPRINTS USING POSITIONS OF SALIENT REGIONS

Chahid Ouali[1,2], Pierre Dumouchel[1] and Vishwa Gupta[2]

[1] ÉTS (École de Technologie Supérieure, Montreal), Canada
[2] CRIM (Computer Research Institute of Montreal), Montreal, Canada
{Chahid.Ouali, Vishwa.Gupta}@crim.ca, Pierre.Dumouchel@etsmtl.ca

ABSTRACT

This paper describes a video fingerprinting system that is highly robust to audio and video transformations. The proposed system adapts a robust audio fingerprint extraction approach to video fingerprinting. The audio fingerprinting system converts the spectrogram into binary images, and then encodes the positions of salient regions selected from each binary image. Visual features are extracted in a similar way from the video images. We propose two visual fingerprint generation methods where fingerprints encode the positions of salient regions of greyscale video images. Salient regions of the first method are selected based on the intensity values of the image, while the second method identifies the regions that represent the highest variations between two successive images. The similarity between two fingerprints is defined as the intersection between their elements. The search algorithm is speeded up by an efficient implementation on a Graphics Processing Unit (GPU). We evaluate the performance of the proposed video system on TRECVID 2009 and 2010 datasets, and we show that this system achieves promising results and outperforms other state-of-the-art video copy detection methods for queries that do not includes geometric transformations. In addition, we show the effectiveness of this system for a challenging audio+video copy detection task.

Index Terms— Content-based copy detection, feature extraction, video fingerprint, fingerprinting.

1. INTRODUCTION

According to a study by the International Data Corporation (IDC), the digital universe is doubling in size every two years to reach 44 trillion gigabytes by 2020 [1]. A large part of this big universe consists of audio and video, which are distributed over the Internet in an effortless way. In fact, video hosting services have facilitated sharing and distributing of video content. For example, 300 hours of videos are uploaded to YouTube every single minute [2]. Unavoidably, a large number of the uploaded videos are illegal copies of digital material protected by copyright law. Copyright infringement is one of the biggest issues that hosting web sites have to deal with to avoid lawsuits by the copyright holders.

To prevent copyright infringements, Content-Based Copy Detection (CBCD) has been recently introduced as an alternative to the watermarking approaches [3]. Instead of inserting additional information to the content, CBCD uses the content itself as a watermark. It extracts relevant features (fingerprints) from a candidate copy and then compares them against fingerprints of the original content. However, audio and video signals are subjected to various kinds of transformations that make the task challenging.

Color-based fingerprints are among the first video features used in video copy detection [4-7]. Despite their popularity, these features are sensitive to several video transformations such as insertion of logos, compression and change of color [6].

The ordinal measure [8] divides each image into N blocks and sort them according to their average grey level. The ordinal measure of a given frame is defined by a vector containing the rank of each block. This technique is used in a similar way in [9]. However, instead of ranking regions in the image, a temporal window is used to rank regions along the time.

Another global feature scheme consists of using a Bag-of-global visual features based on a DCT-sign-based feature [10]. They performed multiple assignments of visual words in the feature, spatial, and temporal domain to improve repeatability of Visual-Words based feature matching. In [11], TIRI-DCT is proposed to generate spatio-temporal fingerprints based on the Temporally Informative Representative Images (TIRI) [12].

In [13], the video is modeled using a graph that represents the relations among different frames. This graph-based modeling scheme achieved good results when evaluated on a small dataset, and performed as well as the TIRI-based method [11]. On the other hand, several papers propose to generate fingerprints based on local information of the image [14-17]. A comparison between global and local features shows that local features outperform global features when evaluated on three different datasets [18]. Similarly in [19], local features based on SIFT show their robustness against transformations that change the content of the video frame compared to the global based feature.

A good multimodal feature representation that exploits the complementary audio features, local visual features and global visual features is described in [20]. The audio part of this system is based on the Weighted Audio Spectrum Flatness (WASF) features introduced in [21]. A local visual feature of dense color SIFT (DC-SIFT) [22] is used as local feature, whereas the global visual feature is based on DCT feature. The similarity search is performed using a temporal pyramid-matching algorithm, where several techniques are employed to speed up the search. This system achieved excellent results on TRECVID 2009 and 2010 datasets when the results obtained by the individual features are combined using a result-level fusion mechanism [20].

In [23], audio fingerprints [24] are used with the ordinal signature [25] to perform a two-step search. First, a number of candidate videos are selected based on the audio results. Then visual features are extracted from the selected candidates and combined with the audio fingerprints to produce the final results.

In this paper we extend our work on multimedia copy detection [26]. The idea behind the proposed video extraction method is adopted from the audio feature extraction method introduced in [27]. Audio fingerprints generated with this method encode the

positions of salient regions of binary images derived from the spectrogram. These spectrogram-based fingerprints have proved their robustness against a variety of audio transformations. In order to see if the audio fingerprint extraction scheme will work well for video fingerprints, we introduce two new visual feature extraction methods and we compare them to two other visual features. Results of this comparison on TRECVID 2009 and 2010 datasets show the robustness of the proposed visual features, especially for queries that do not include geometric transformations. We also describe a simple fusion technique to detect video queries transformed by audio and video transformations. We show that the overall system achieves excellent video detection performance for the task of audio+video copy detection.

2. SYSTEM OVERVIEW

First, we convert each reference video into a sequence of greyscale images, and we change their sizes to a fixed size (width = height = 300 pixels). After preprocessing, we extract fingerprints from these images and we store them into a video fingerprints database.

Video queries go through many complex transformations and may contain a combination of transformations. Hence, we generate fingerprints for the original video query and a flipped version of the video query. Besides, we propose a Picture in Picture (PiP) detection algorithm that detects PiP from the query and then extracts the foreground video before extracting the features.

2.1. PiP Detection

First, we select a fixed number of images from the video query (in our experiments we select 50 images). These images are selected uniformly and regardless of the video length. We process in this way to avoid handling all the query frames and reduce, therefore, the processing time. Then, we divide each selected image into 5 regions (four corners and the center), where the size of each region is equal to the half of the original image size. Then, each selected image region is processed as described in Figure 1.

1. For each region of the selected images:
    2. Perform edge detection
    3. Perform Hough transform and extract line segments (only horizontal and vertical lines)
    4. Detect locations (pixel positions) of perpendicular lines
    5. Increment the number of the detected locations
6. Merge closest locations
7. Get the top-4 locations (number of apparitions)
8. Verify if these locations form a rectangle

Figure 1. PiP detection steps.

Figure 2 shows results of performing steps 2-4 of Figure 1 on six images (central region of the original image), where intersection points of the detected segment lines are marked with red points. Notice that we keep only horizontal and vertical line segments, and we extend the extremities of each segment to force intersection of short segments. For every processed image, we keep locations of intersection points and increment the number (score) of their appearances (step 4-5). Once all the images have been processed, we merge locations and scores of points that are very close (step 6), and keep the four most frequent intersection points (step 7). These top-4 points represent corners of the candidate PiP region. Finally in step 8, we verify if these corners form a rectangle based on some criteria (size of sides, parallel and vertical sides).

Figure 2. Example of PiP detection.

2.2. Video Fingerprint Extraction

The proposed feature extraction is adopted from the audio feature extraction method described in [27]. This method generates binary images from the spectrogram and then encodes the position of several salient regions that have the highest spectral values in each binary image. Video feature extraction is based on the same idea as the audio feature extraction: positions of salient regions of an image have good chance to survive signal degradation. The question is how can we define a visual salient region?

Audio fingerprints are extracted from a binary image derived from the spectrogram, where a value of 1 denotes a time-frequency peak. In other words, a value of 1 indicates the presence of information, and a value of 0 denotes absence of information. Thus, a salient region is the part of the binary image that has more information than the others. Multiple audio fingerprints describe, therefore, the localization of information over time regardless of the real intensity values. In contrary, video images are more complex, and each pixel may hold useful information.

We propose two different fingerprint extraction schemes: V-intensity and V-motion fingerprints. For these two extraction methods, we divide the image using a tile of size 20 × 20 for a total of 225 squares, and we compute the sum of the pixel values in each square. Then, each method selects d squares per image. The positions of the selected squares (i.e. salient regions) represent the final fingerprint. V-motion and V-intensity fingerprints are selected as follows:

V-intensity: this method sorts the squares by their values and takes $d/2$ squares before and $d/2$ squares after the square with the median value. In other words, we take image regions that are neither black nor white, but grey regions. The grey regions are the regions of interest for distinguishing video frames. An illustration of this method is given in Figure 3, where the selected regions are represented in grey background.

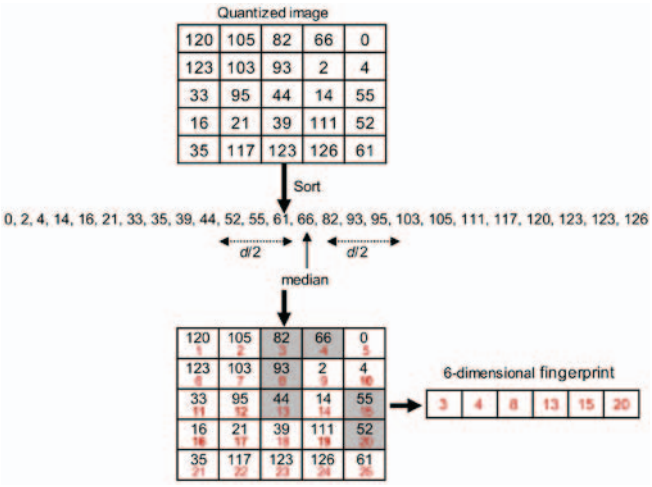

Figure 3. Illustration of V-intensity fingerprint generation scheme.

V-motion: V-motion fingerprints are detected by looking for the regions that have the highest variations compared to the same regions in the previous frame. Figure 4 illustrates the principal steps to extract V-motion fingerprints. In this figure, the square values difference between frame 1 and frame 2 indicates the degree of intensity variations between these two successive frames. The regions that have the highest variations are the salient regions.

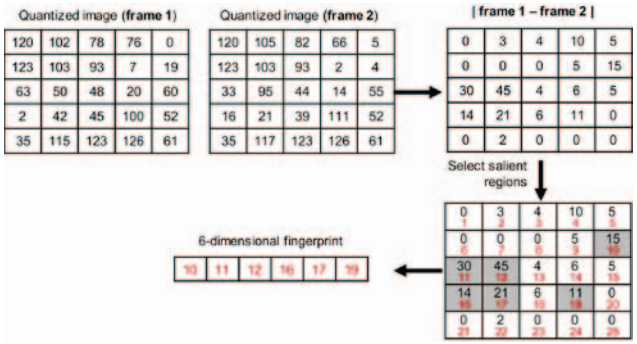

Figure 4. Illustration of V-motion fingerprint generation scheme.

2.4. Fingerprint retrieval

The fingerprint retrieval algorithm is composed of two principal steps: a similarity search algorithm followed by a matching step. The similarity search labels each reference frame with the closest query frame. In the second step, we move the query over the references to compute the number of matching frames between them. A detailed description of this algorithm can be found in [27].

Computing the distance between each reference fingerprint and all query fingerprints is time consuming. This is a common problem in the CBCD task, where tens of millions of fingerprints are generated from multimedia dataset. In order to accelerate the similarity search algorithm, we implemented the similarity search algorithm on a Graphics Processing Unit (GPU). Compared to the CPU implementation, the GPU implementation accelerated the similarity search by over 150 times [28].

3. EXPERIMENTS

This section evaluates our system on TRECVID 2009 and 2010 copy detection datasets. First, we evaluate our system for the video only task, while comparing the results to two video fingerprinting methods. Then, we give results for the audio+video copy detection task (i.e. the query is transformed by audio and video transformations). Finally, we compare the audio+video results with the best published results on TRECVID 2010 dataset.

Table 1. Description of audio and video transformations.

Type	Label	Description
Audio transformation	T1	Nothing
	T2	mp3 compression
	T3	mp3 compression and multiband companding
	T4	bandwidth limit and single band companding
	T5	mix with speech
	T6	mix with speech, then multiband compress
	T7	bandpass filter, mix with speech, compress
Video transformation	V1	Simulated camcording
	V2	Picture in picture type 1: original video in front of background video
	V3	Insertions of pattern
	V4	Strong re-encoding
	V5	Change of gamma
	V6	Decrease in quality: introducing 3 randomly selected combination of Blur, Gamma, Frame dropping, Contrast, Compression, Ratio, White noise
	V8	Post production: introducing 3 randomly selected combination of Crop, Shift, Contrast, Text insertion, Vertical mirroring, Insertion of pattern, Picture in picture
	V10	Combination of 3 randomly selected transformations chosen from V1-V8

3.1. Datasets

TRECVID 2009 and 2010 datasets are provided by NIST [29]. Each of these datasets consists of a reference collection of about 400 hours of videos. There are 201 original queries, each query altered with 7 audio transformations (for a total of 1407 audio queries) and 8 video transformations (for a total of 1608 video queries). Descriptions of audio and video transformations are shown in Table 1. The total number of audio+video queries for TRECVID 2010 is equal to 11256 containing 56 transformations (7 audio × 8 video transformations). In TRECVID 2009, 49 audio+video transformations (7 audio × 7 video transformations) results in 9849 audio+video queries. Note that V1 transformation is not applied in TRECVID 2009. To evaluate the copy detection performance, we use the minimal Normalized Detection Cost Rate (min NDCR). NDCR is a weighted cost combination of the probability of missing a true copy and the false alarm rate.

3.2. Video Only Results

The experimental results on TRECVID 2009 and TRECVID 2010 datasets for V-intensity and V-motion features are shown in Table 2. These results are compared with DC-SIFT and DCT features. From Table 2 it can be seen that V-motion performs better than V-intensity for all transformations and on both datasets (except V4 on TRECVID 2009). Although these two features achieved good results for transformations V3, V4, V5 and V6 that do not include geometric transformations, they give relatively higher min NDCR for transformations that change the content of the images. This is

noticeable for transformation V1 that gives the highest min NDCR compared to the rest of the transformations. In fact, global visual features are usually sensitive to such transformations, as confirmed by the results of DCT feature.

Table 2. Min NDCR per transformation achieved by different visual features on TRECVID 2009 and 2010 datasets.

	Feature	V1	V2	V3	V4	V5	V6	V8	V10
2009	V-intensity	-	0.351	0.56	0.007	0.007	0.007	0.552	0.448
	V-motion	-	0.284	0	0.045	0	0	0.231	0.313
	DC-SIFT	-	0.112	0.03	0.09	0.024	0.142	0.201	0.149
	DCT	-	0.224	0.164	0.119	0.104	0.231	0.41	0.306
2010	V-intensity	0.985	0.634	0.276	0.097	0.067	0.149	0.53	0.463
	V-motion	0.896	0.545	0.03	0.082	0.03	0.112	0.321	0.358
	DC-SIFT	0.285	0.154	0.054	0.146	0.038	0.223	0.292	0.2
	DCT	1	0.377	0.246	0.2	0.146	0.323	0.585	0.415

The comparison between DCT and V-motion shows that V-motion gives better results for most of the transformations on both the datasets. DC-SIFT works better on transformations that modify the content and achieved the best results for V1, V2 and V10 transformations on TRECVID 2009. In fact, V-motion gave the best results for four transformations (V3, V4, V5 and V6) on both datasets. For these four transformations, V-motion missed only one query (transformed with V4) on TRECVID 2009 (see Table 3).

Table 3. Number of missed queries for V-intensity and V-motion on TRECVID 2009 dataset.

Feature	V1	V2	V3	V4	V5	V6	V8	V10
V-Intensity	-	45	14	1	1	0	65	57
V-motion	-	36	0	1	0	0	23	35

In order to evaluate the performance of PiP detection, we count the number of PiPs correctly detected for both datasets in Table 4. As mentioned above, there were 201 PiP queries composed of 134 reference copies and 67 queries from non-reference videos (i.e. false alarms). As can be seen from Table 4, the PiP algorithm detects between 73% and 79% of the inserted PiPs for TRECVID 2009 and 2010 datasets, respectively.

Table 4. PiP detection performance.

Dataset	Detected	Missed	% of detection
TRECVID 2009	98	36	73%
TRECVID 2010	106	28	79%

3.3. Audio+Video Results

We use a simple strategy to combine audio and video results. First, we generate results separately for the audio [27] and the proposed video systems. Then for each query, we keep the best result (highest score) achieved by either the audio or the video. In other words, for a given query, if the best reference audio score is higher than the best video reference score, then we take the audio result, otherwise, we take the video result. This fusion results in a very good performance. In fact, the min NDCR averaged over all transformations is equal to 0.021 and 0.053 for TRECVID 2009 and 2010 datasets, respectively. From a total of 9849 queries in TRECVID 2009, our system missed only 122 queries (98.7% correctly detected). On TRECVID 2010, 347 queries are missed from a total of 11256 queries (96.9 % correctly detected).

Finally, we compare in Figure 5 our audio+video system to the method described in [20] (the Perseus system) that combines the results obtained by the audio part using WASF feature, and the video results obtained using DC-SIFT and DCT visual features. This method achieved the best results for almost all the transformations in the TRECVID 2010 evaluation campaign [30].

It can be seen from Figure 5 that our system achieved comparable results to the Perseus system and outperforms it for 35 out of 56 transformations. Furthermore, our system gave a lower min NDCR (averaged over all transformations) of 0.056 compared to 0.06 achieved by Perseus system.

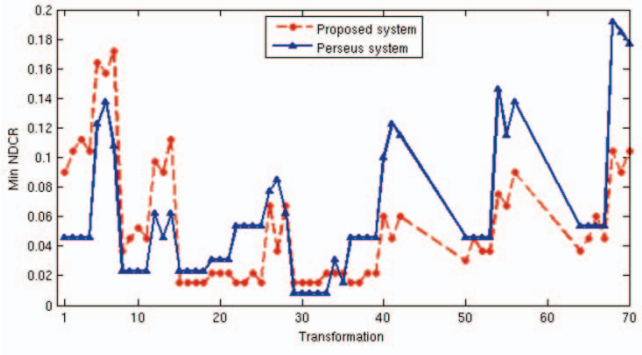

Figure 5. Min NDCR of the proposed system for audio+video transformations compared to Perseus system on TRECVID 2010.

4. CONCLUSION

This paper describes a robust multimedia fingerprinting system that can be used to detect video copies subjected to complicated audio and video transformations. The proposed video feature extraction is similar to a state-of-the-art audio copy detection feature extraction strategy that converts the audio into a set of binary images, and then encodes the positions of several selected regions from each binary image. In this work, instead of extracting fingerprints from binary images, visual features are extracted from greyscale video images (for robustness to color transformations).

We propose V-intensity and V-motion features and show that V-motion is more robust to video transformations than V-intensity. V-intensity selects salient regions based on the intensity values of the image, while V-motion identifies the regions that represent the highest intensity variations between two successive images. We compare these two methods to DCT and DC-SIFT features using TRECVID 2009 and 2010 datasets. We show that V-motion achieves excellent results for all queries that do not include geometric transformations and outperforms the other features for these transformations. To address the PiP transformation, we propose a PiP detection technique that detects 79% of PiP on TRECVID 2010 dataset. We also tested our system for the audio+video copy detection task where the queries are transformed by a combination of audio and video transformations. These queries are detected by taking the best result achieved by the audio and video systems. This audio+video merging works very well and it gave excellent min NDCR of 0.021 and 0.053 for TRECVID 2009 and 2010 datasets, respectively. This compares well with the best published min NDCR of 0.06 for the TRECVID 2010 dataset for the no false alarm case.

5. REFERENCES

[1] V. Turner, J. F. Gantz, D. Reinsel *et al.*, “The digital universe of opportunities: Rich data and the increasing value of the internet of things,” *International Data Corporation, White Paper, IDC_1672*, 2014.

[2] Youtube. "YouTube Statistics," 20 mars 2015; https://www.youtube.com/yt/press/statistics.html.

[3] F. Hartung, and M. Kutter, “Multimedia watermarking techniques,” *Proceedings of the IEEE,* vol. 87, no. 7, pp. 1079-1107, 1999.

[4] J. Law-To, L. Chen, A. Joly *et al.*, "Video copy detection: A comparative study," *Proceedings of the 6th ACM International Conference on Image and Video Retrieval, CIVR 2007.* pp. 371-378.

[5] M. R. Naphade, M. M. Yeung, and B.-L. Yeo, “Novel scheme for fast and efficent video sequence matching using compact signatures,” in Electronic Imaging, 1999, pp. 564-572.

[6] H. T. Shen, X. Zhou, Z. Huang *et al.*, “UQLIPS: a real-time near-duplicate video clip detection system,” in Proceedings of the 33rd international conference on Very large data bases, 2007, pp. 1374-1377.

[7] X. Wu, A. G. Hauptmann, and C.-W. Ngo, “Practical elimination of near-duplicates from Web video search,” in 15th ACM International Conference on Multimedia, MM'07, September 24, 2007 - September 29, 2007, Augsburg, Bavaria, Germany, 2007, pp. 218-227.

[8] A. Hampapur, K.-H. Hyun, and R. Bolle, “Comparison of sequence matching techniques for video copy detection,” in Storage and Retrieval for Media Databases 2002, January 23, 2002 - January 25, 2002, San Jose, CA, United states, 2002, pp. 194-201.

[9] L. Chen, and F. W. M. Stentiford, “Video sequence matching based on temporal ordinal measurement,” *Pattern Recognition Letters,* vol. 29, no. 13, pp. 1824-1831, 2008.

[10] Y. Uchida, K. Takagi, and S. Sakazawa, "KDDI labs at TRECVID 2011: Content-based copy detection," *2011 TREC Video Retrieval Evaluation Notebook Papers.* p. National Institute of Standards and Technology (NIST).

[11] M. M. Esmaeili, M. Fatourechi, and R. K. Ward, “A robust and fast video copy detection system using content-based fingerprinting,” *IEEE Transactions on Information Forensics and Security,* vol. 6, no. 1, pp. 213-226, 2011.

[12] M. Malekesmaeili, M. Fatourechi, and R. K. Ward, "Video copy detection using temporally informative representative images," *8th International Conference on Machine Learning and Applications, ICMLA 2009.* pp. 69-74.

[13] X. Nie, W. Zeng, H. Yan *et al.*, “Structural similarity-based video fingerprinting for video copy detection,” *IET Image Processing,* vol. 8, no. 11, pp. 655-661, 2014.

[14] M.-C. Yeh, C.-Y. Hsu, and C.-S. Lu, "NTNU-Academia Sinica at TRECVID 2010 content based copy detection," *2010 TREC Video Retrieval Evaluation Notebook Papers.* p. National Institute of Standards and Technology (NIST).

[15] M. Douze, H. Jegou, and C. Schmid, “An image-based approach to video copy detection with spatio-temporal post-filtering,” *IEEE Transactions on Multimedia,* vol. 12, no. 4, pp. 257-266, 2010.

[16] M. Heritier, V. Gupta, L. Gagnon *et al.*, "Crim's content-based copy detection system for TRECVID," *2009 TREC Video Retrieval Evaluation Notebook Papers.* p. National Institute of Standards and Technology (NIST).

[17] Z. Liu, T. Liu, and B. Shahraray, "ATT research at TRECVID 2009 content-based copy detection," *2009 TREC Video Retrieval Evaluation Notebook Papers.* p. National Institute of Standards and Technology (NIST).

[18] C.-L. Chou, H.-T. Chen, and S.-Y. Lee, “Pattern-Based Near-Duplicate Video Retrieval and Localization on Web-Scale Videos,” *Multimedia, IEEE Transactions on,* vol. 17, no. 3, pp. 382-395, 2015.

[19] C.-Y. Chiu, T.-H. Tsai, Y.-C. Liou *et al.*, “Near-duplicate subsequence matching between the continuous stream and large video dataset,” *Multimedia, IEEE Transactions on,* vol. 16, no. 7, pp. 1952-1962, 2014.

[20] L. Mou, T. Huang, Y. Tian *et al.*, “Content-based copy detection through multimodal feature representation and temporal pyramid matching,” *ACM Transactions on Multimedia Computing, Communications and Applications,* vol. 10, no. 1, 2013.

[21] J. Chen, and T. Huang, “A robust feature extraction algorithm for audio fingerprinting,” in 9th Pacific Rim Conference on Multimedia, PCM 2008, December 9, 2008 - December 13, 2008, Tainan, Taiwan, 2008, pp. 887-890.

[22] A. Bosch, A. Zisserman, and X. Munoz, “Scene classification using a hybrid generative/discriminative approach,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 30, no. 4, pp. 712-727, 2008.

[23] T. Li, F. Nian, X. Wu *et al.*, “Efficient video copy detection using multi-modality and dynamic path search,” *Multimedia Systems*, pp. 1-11, 2014.

[24] J. Haitsma, and T. Kalker, “A Highly Robust Audio Fingerprinting System,” in Ismir, 2002, pp. 107-115..

[25] D. N. Bhat, and S. K. Nayar, “Ordinal measures for image correspondence,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 20, no. 4, pp. 415-423, 1998.

[26] C. Ouali, P. Dumouchel, and V. Gupta, “Content-Based Multimedia Copy Detection,” in IEEE International Symposium on Multimedia, Miami, Florida USA, pp. 597-600. ,2015.

[27] C. Ouali, P. Dumouchel, and V. Gupta, “Efficient Spectrogram-Based Binary Image Feature For Audio Copy Detection,” in 40th IEEE International Conference on Acoustics, Speech and Signal Processing, Australia, pp. 1792-1796, 2015.

[28] C. Ouali, P. Dumouchel, and V. Gupta, “GPU Implementation of an Audio Fingerprints Similarity Search Algorithm,” in Content-Based Multimedia Indexing, Prague, pp. 1-6, 2015.

[29] G. Awad, P. Over, and W. Kraaij, “Content-based video copy detection benchmarking at TRECVID,” *ACM Transactions on Information Systems (TOIS),* vol. 32, no. 3, pp. 14, 2014.

[30] Y. Li, L. Mou, M. Jiang *et al.*, “PKU-IDM@ TRECVid 2010: copy detection with visual-audio feature fusion and sequential pyramid matching,” *online] wwwnlpir. nist. gov/projects/tvpubs/tv. pubs. org. html*, 2010.