CONTENTs

Message from MMTC Chair ................................................................. 3

SPECIAL ISSUE ON LARGE-SCALE MULTIMEDIA COPY DETECTION
ALGORITHMS AND SYSTEMS .......................................................... 4

Large-scale Multimedia Copy Detection Algorithms and Systems ........................................ 4
  Guest Editor: Zhu Liu, AT&T, USA .................................................. 4
  zliu@research.att.com ................................................................. 4

Multimodal Video Copy Detection using Multi-Detectors Fusion ..................................... 6
  Yonghong Tian, Tiejun Huang, and Wen Gao (Fellow, IEEE) ........................................ 6
  {yhtian, tjhuang, wgao}@pku.edu.cn ............................................ 6

Near-Duplicate Image Retrieval as a Classification Problem ........................................ 10
  Yusuke Uchida and Shigeuki Sakazawa ........................................ 10
  KDDI R&D Laboratories, Inc., Japan ............................................ 10
  {ys-uchida, sakazawa}@kddilabs.jp ............................................ 10

Robust Media Search Technology for Content-Based Audio and Video Identification ...... 13
  Ryo Mukai, Takayuki Kurozumi, Takahito Kawanishi, Hidehisa Nagano .................... 13
  and Kunio Kashino ................................................................ 13
  NTT Corporation, Japan ............................................................ 13
  {mukai.ryo, kurozumi.takayuki, kawanishi.takahito, nagano.hidehisa,
  kashino.kunio}@lab.ntt.co.jp ..................................................... 13

Multimodal Video Copy Detection using Local Features ................................................ 17
  Xavier Anguera1 and Tomasz Adamek2 ........................................ 17
  1Telefónica Research, Spain ....................................................... 17
  2Catchoom, Spain ................................................................ 17
  xanguera@tid.es .................................................................. 17

Datasets for Content-Based Video Copy Detection Research ........................................ 20
  Yudong Jiang and Yu-Gang Jiang ................................................ 20
  Fudan University, China ............................................................ 20
  {11210240091, ygj}@fudan.edu.cn ........................................... 20

Large-scale Multimedia Content Copy Detection in Mainstream Applications ............... 23
  Eric Zavesky ........................................................................ 23
  AT&T Labs Research, USA ....................................................... 23
1. Introduction

Content-based copy detection (CBCD) is drawing increasing attention for content-based retrieval, media usage monitoring, content identification and copyright protection [1]. For example, copy detection can be used to prevent users from uploading copyrighted material to YouTube, who ever claimed in April 2012 that a staggering 72 hours of video were uploaded to the site every minute and meanwhile was harshly criticized for failing to ensure that these uploaded videos comply with the law of copyright.

Technologically, however, copy detection in a Web-scale video database is a pretty challenging task. This is mainly due to the fact that Web video copies often suffer different complex transformations on the components of audio, video, or both of a video file or stream, and to make things worse, the content of many copies may be significantly changed from their originals. After years of practice, it has been widely recognized that none of any single feature, or single detector based on several features, can work well for all transformations. Thus, it is beneficial to combine multimodal features or several detectors to enhance the robustness and discriminability of the copy detection system. This trend has also been validated by recent practices in the TRECVID-CBCD contests [2], in which most of the participating approaches compute several detection results through individual features and then fuse them to obtain the final result.

In this article, we briefly describe our approaches that achieved the best copy detection accuracies and excellent localization precisions for the majority of transformations at 2010 and 2011 TRECVID-CBCD contests. Our basic idea is to exploit complementary audio-visual features to construct several detectors and then organize them in a principled structure. Here each detector often adopts the frame fusion based paradigm [11], namely, searching a list of similar reference frames for each query frame and then determining video-level matches from these frame-level matches. Given video-level matches from these detectors, two multi-detectors fusion methods namely, verification-based fusion and CBCD-oriented detector cascading, are designed to derive the final detection result. In the following, we will describe them one by one.

2. Verification-based Multimodal Fusion

In the first approach, we employ a verification-based fusion method to combine the detection results from different detectors. As shown in Figure 1, four multimodal features are used to construct detectors, including two local visual features SIFT [3] and SURF [4], a global visual feature based on DCT and an audio feature WASF [5]. Comparatively, local visual features can effectively handle spatial content-altering transformations (e.g., cropping, picture-in-picture (PiP), pattern insertion) while global visual features are capable of resisting spatial content-preserving but quality-degrading operations (e.g., noise addition, resolution change, re-encoding). The two local visual features are used here mainly because a copy that is asserted as a non-copy by one feature might be detected as a copy by the other. To speed up feature searching, bag-of-words (BoW) technique is applied to convert each SIFT or SURF descriptor into a visual word (400 words generated from 2M descriptors) and then the inverted index is used for indexing the BoWs of SIFT and SURF. Meanwhile, local sensitive hashing (LSH) [6] is also used to accelerate similarity search for descriptors of DCT and WASF.

For frame fusion based paradigm, another key issue is to define appropriate temporal constraints on frame-level matches such that two matched sequences have consistent timestamps. To address this problem, we proposed temporal pyramid matching (TPM) in [7]. Inspired by spatial pyramid matching [8] which conducts pyramid match kernel in 2-D image space, TPM partitions each video into increasingly finer temporal segments and assemble frame-level similarity search results into video-level matches through similarity evaluation on multiple temporal granularities.

Considering that the BoW representation inevitably causes decrease in feature’s discriminability, a verification mechanism is added in the result-level fusion module. That is, if a query video is simultaneously asserted by at least two detectors as a copy of the same reference video, then it is accepted as a copy; otherwise, if a query video is reported as a copy only by one detector, it should be further verified using the original SIFT descriptor. In this case, only if the recalculated similarity for the video-level match is above a pre-defined threshold, will it be accepted as a real copy.
It should also be noted that in our system, PiP is detected by Hough transform which detects two pairs of parallel lines so as to locate the inserted foreground videos. For those queries with PiP, the foreground and original key-frames will be processed respectively. In addition, queries asserted as non-copies will be flipped and matched again to deal with flip transformation.

We submitted two runs to the 2010 TRECVID-CBCD contest. Official evaluation results show that among totally 56 transformations, our system achieved excellent NDCR (Normalized Detection Cost Rate) \cite{8}: 39 best “Actual NDCR” and 51 best “Optimal NDCR” for BALANCED profile; 52 best “Actual NDCR” and 50 best “Optimal NDCR” for NOFA profile. For localization precision, our system also obtained competitive F1: averagely 0.9 for both BALANCED and NOFA profiles and all the transformations. Nevertheless, such an excellent detection performance was obtained at the cost of a long processing time, despite the system could be optimized with multi-thread programming.

3. CBCD-oriented Detector Cascading

To solve the efficiency issue, we proposed a copy detection approach with a cascade of multimodal features. In this CBCD-oriented cascade architecture (See Figure 2), detectors based on several complementary audio-visual features are organized in a cascade structure such that efficient but relatively simple detectors are placed in the front, while effective but complex detectors are located in the rear. Thus with elaborately tuned decision thresholds for these detectors, the processing time can be significantly reduced for most copies since they can be correctly detected through at most the first two detectors.

Instead of using both SIFT and SURF in our previous system at TRECVID-CBCD 2010, here we only use only one local feature, i.e., DC-SIFT \cite{9}, since DC-SIFT can better represent scenes as well as objects, leading to a better performance in video copy detection task. More importantly, inspired by the well-known “classifier cascade,” our system places a series of detectors in a simple-to-complex order. Formally, in an $N$-Stage cascade of detectors, $D_q = \langle d_1, d_2, \ldots, d_N \rangle$, a query $q$ is processed by each detector successively until one asserts it as a copy or all determine it as a non-copy. That is, $q$ is first processed by $d_1$ where a positive detection result, i.e., the returned reference video $r_1$ has a similarity $s_1^{(V)}$ greater than or equal to a predefined threshold $\theta_1$, leading to the immediate acceptance of $q$ as a copy; otherwise, the evaluation of $d_2$ on $q$ will be triggered... Only if $q$ is asserted as a non-copy by all the detectors, will it be accepted as a non-copy. In practice, most copies can be detected through the first two detectors, thus saving a major part of processing time.

The effectiveness of our CBCD-oriented cascading architecture can be further illustrated by Table 1. We can see that only the WASF detector is enough to deal with A1-A4 audio transformations, no matter which visual transformations exist; otherwise, if one of V3-V6 visual transformations is also exerted on the query video, two detectors respectively over WASF and DCT are needed; for the remaining cases, all three detectors are needed. Our experimental results also validate the complementarity of the individual detectors in our system.

Figure 1. Framework of our copy detection system used for 2010 TRECVID-CBCD task

http://www.comsoc.org/~mmc/
Official evaluation results in the TRECVID-CBCD 2011 contest showed that our system also achieved excellent NDCR performance (i.e., 34 best “Actual NDCR” and 31 best “Optimal NDCR” for BALANCED profile; 31 best “Actual NDCR” and 14 best “Optimal NDCR” for NOFA profile) and very good F1 performance (i.e., average F1 of 0.95 for both BALANCED and NOFA profiles and all the transformations). More importantly, due to the adoption of CBCD-oriented cascade architecture, our Processing Times were shorter than the median ones of all the participants, and also much less that our previous system at TRECVID-CBCD 2010.

Table 1. The effectiveness of different features on the audio-visual transformations.

<table>
<thead>
<tr>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
<th>A6</th>
<th>A7</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V5</td>
<td></td>
<td></td>
<td></td>
<td>Case3: WASF + DCT + DC-SIFT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V10</td>
<td></td>
<td>Case2: WASF + DCT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3 shows the performance curves of different methods over 56 transformations for BALANCED profile, including our system (“Our-D3”), its simplified version only with WASF and DCT detectors (“Our-D2”), two best approaches from other 21 participants (CRIM-VISI and INRIA-LEAR), and the median performance on each transformation among all approaches (“Median”). We can see that when using only WASF and DCT detectors, Our-D2 could obtain a slightly less excellent NDCR and better F1 results with a small fraction of processing time than Our-D3.

4. Summary
This article presents two multimodal video copy detection approaches that exploit complementary audio-visual features to detect copies that are subjected...
to complicated transformations. Due to the scalable processing performance, our approaches are capable of satisfying various requirements in many Web-scale applications.

References


Yonghong Tian is currently an associate professor at School of EE & CS, Peking University, China. His research interests include computer vision and multimedia content analysis. He has published over 70 technical papers, and serves on the OC and TPC of several conferences such as ACM MM 2009, KDD 2010, and MMSP 2011. His team has been the top performer in the WikepediaMM task at ImageCLEF2008 and 2010/2011 TRECVID-CBCD task, and one of the top performers in the 2009/2010/2011 TRECVID-SED task. He is a senior member of IEEE.

Tiejun Huang is currently a professor at School of EE & CS, Peking University and the deputy director of the National Engineering Laboratory of Video Technology, Beijing, China. His research interests include image understanding, video coding, and digital rights management. He has published over 90 technical papers. He is a Board member of Digital Media Project and an advisory Board of IEEE Computing Now.

Wen Gao is currently a professor at School of EE & CS, Peking University, China. He has published four books and over 600 technical articles in refereed journals and proceedings in the areas of signal processing, video communication, computer vision, multimodal interface, and pattern recognition. His current research interests include all fields of digital media technology. He has served the academic community in many positions, including as the General Co-Chair of IEEE ICME 2007, ACM MM 2009, IEEE MMSP 2011, IEEE ISCAS 2013. He is a fellow of IEEE and Academician of Chinese Academy of Engineering.