MULTIMEDIA COMMUNICATIONS TECHNICAL COMMITTEE
IEEE COMMUNICATIONS SOCIETY
http://www.comsoc.org/~mmc

E-LETTER

Vol. 7, No. 7, September 2012

IEEE COMMUNICATIONS SOCIETY

CONTENTS

Message from MMTC Chair

Dear MMTC fellow members,

Welcome to the September issue of IEEE MMTC E-letter! Time really flies. This is the first time I write to you as the vice-chair of MMTC. First of all, I want to express my sincere gratitude to the E-letter and R-letter editorial teams as well as chairs and vice–chairs of special interest groups (IG), and all contributors over the past two years. Their contribution and dedication have made both E-letter and R-letter valuable platforms for researchers in the multimedia communications society to share and exchange ideas. Special thanks also go to the former and current MMTC chairs, Dr. Haohong Wang and Dr. Jianwei Huang for their great leadership and strong support to E-letter and R-letter. Thank you!

Looking forward, E-letter and R-letter editorial teams will continue to work with special issue groups of MMTC to bring to MMTC members the latest technological developments, hot research topics, and society news in multimedia communications and signal processing. Please follow closely our E-letter and R-letter.

It is with my great pleasure to announce the new E-letter leadership team for the term of 2012-2014:

- E-letter Director, Dr. Shiwen Mao (Auburn University, USA)
- E-letter Co-director, Dr. Guosen Yue (NEC Labs, USA)
- E-letter Co-director, Dr. Periklis Chatzimisios (Alexander Technological Educational Institute of Thessaloniki, Greece)

Dr. Mao has served as MMTC IG Chair during the past two years. Starting from July 2012, E-letter will resume to be a bimonthly publication. Each issue of E-letter will feature a special issue on emerging topics that contributed by an IG and E-letter jointly and an industry column organized by an E-letter editor or guest editor. The special issue on emerging topics may consist of two types of papers, i.e., extended abstract from open call for papers or short invited papers from well-known researchers summarizing their recent works, possibly a paper published in premier IEEE publications. The full version of an accepted abstract will be recommended to fast-track review of an IEEE special issue organized by the IG. The industry column will focus on practical industry problems and solutions. Historically, many excellent E-letter contributions are position papers partially based on existing publications in top IEEE ComSoc conferences and journals. In the future, the E-letter board will make selected recommendations of these related existing publications to the R-letter board, further increasing the visibility of the E-letter contributors.

I am also very happy to introduce to you the new R-letter leadership team for the term of 2012-2014:

- R-letter Director, Dr. Irene Cheng (University of Alberta, Canada)
- R-letter Co-director, Dr. Xianbin Wang (The University of Western Ontario, Canada)

Both Dr. Cheng and Dr. Wang have served as MMTC IG Chairs during the past two years. In the next two years, the new R-letter editorial board will continue to stimulate multimedia research by selecting high-quality papers on emerging topics for review. The board will also encourage researchers to submit papers to IEEE MMTC sponsored publications and conferences, and nominate papers published in IEEE MMTC sponsored publications/conferences for best paper awards.

I hope you enjoy reading the September issue of E-letter, and as always, your great suggestions and comments are very welcome!

Kai Yang

Vice-Chair of Multimedia Communication TC of IEEE ComSoc

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

Large-scale Multimedia Copy Detection Algorithms and Systems

Guest Editor: Zhu Liu, AT&T, USA

zliu@research.att.com

Multimedia copy detection is to locate the copied segments of a target video in a testing video, and it is essential for many applications, including copyright infringement tracking, video content search, monitoring, and storage, etc. Various transformations may be applied in the copied content either intentionally (e.g., to elude the copyright infringement) or unintentionally (e.g., to convert it to the right format for streaming). They include embedding the copied video in the Picture-in-picture mode, strong re-encoding, insertion of text and patterns, rotation, color enhancement, etc. Such modification on the original content demands high robustness of the copy detection algorithms. In addition, with the explosive growth in the volume of multimedia content as well as the desire to link content in real time, the underlying copy detection algorithms have to be efficient and scalable.

This special issue of E-Letter focuses on the recent progresses of large-scale multimedia content copy detection algorithms and systems. It is the great honor of the editorial team to have six leading research groups, from both academia and industry laboratories, to report their solutions for meeting these challenges and share their latest results.

In the first article titled, "*Multimodal Video Copy Detection using Multi-Detectors Fusion",* Tian, Huang, and Gao from Peking University presented their video copy detection system that achieved the best overall accuracy and excellent localization precision in the TRECVID 2010 and 2011 content-based copy detection (CBCD) evaluation task hosted by NIST. Complementary audio-visual features are exploited to construct several detectors and they are organized in a principled structure. Two multi-detectors fusion methods are investigated to determine the final detection results based on the video-level matches, namely, verification-based fusion and CBCD-oriented detector cascading. The proposed approaches are scalable and are suitable for many Web-scale applications.

Uchida and Sakazawa from KDDI R&D Laboratories authored the second article, "Near-Duplicate Image Retrieval as a Classification Problem". The authors proposed a scoring method for local visual feature based image retrieval system, where the descriptors can be optionally quantized. The new score is based on the ratio of the probability density functions of an object model and a background model. This method is generally applicable in many image retrieval systems, and its effectiveness has been demonstrated by the simulation results in the bag-of-visual words-based framework and the *k*-NN voting framework.

The third article is contributed by Mukai *et al.* from NTT Communication Science Laboratories, and the title is *"Robust Media Search Technology for Content-Based Audio and Video Identification"*. A robust media search technology is introduced where the media data is converted into sequences of fingerprints and the spatiotemporal accumulation of them provides extremely high identification accuracy. The adopted audio fingerprint technique is the divide-and-locate method, and the video one is the coarsely-quantized area matching method. The core technology has been deployed for many commercial services, including Internet copyright enforcement, media search on smartphones, and media link analysis.

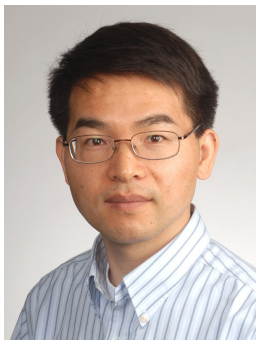
Anguera and Adamek presented a content based copy detection system built by Telefonica Research for NIST's TRECVID evaluation in the fourth article, *"Multimodal Video Copy Detection Using Local Features"*. The system relies on both audio features (MASK) and local video features (DART), and adopts a late fusion algorithm to combine the results from different modalities effectively. It achieved outstanding results in TRECVID 2010 and 2011. A disk-based inverted file index over SSD drives makes the system scalable for large database with low query latency.

The fifth article is *"Datasets for Content-Based Video Copy Detection Research"*, from Jiang and Jiang at Fudan University. Instead of discussing fundamental video copy detection algorithms, the authors provided a nice overview on benchmark datasets widely adopted in this field, which are essential for evaluating different copy detection techniques and encouraging the innovation. In addition, the article introduced a new benchmark dataset that is being created by the group.

The uniqueness of this new dataset is that all videos are downloaded from the Internet without any artificial transformations. It will cover 50 topics, each contains 20 videos. Once it is available in early 2013, it will make significant contribution to the community since researchers can test the algorithms on a collection of real data.

The last article of this special issue is from Zavesky at AT&T Labs – Research, and the title is *"Large-scale Multimedia Content Copy Detection in Mainstream Applications"*. This article provides an excellent high level overview on the applications of video copy detection algorithms. The author first discussed three interesting topics: consumer media management, blind source reconstruction, product & companion linking, and reported AT&T's activities in each of these three areas. Then a brief review on active research and evaluation is given, followed by an insight on the foreseeable challenges in future research.

While this special issue is far from delivering a complete coverage on this exciting research area, we hope that the six invited letters give the audiences a taste of the main activities in this area, and provide them an opportunity to explore and collaborate in the related fields. Finally, we would like to thank all the authors for their great contribution and the E-Letter Board for making this special issue possible.



Zhu Liu received the B.S. and M.S. degrees in Electronic Engineering from Tsinghua University, Beijing, China, in 1994 and 1996, respectively, and the Ph.D. degree in Electrical Engineering from Polytechnic Institute of New York University, Brooklyn, NY, in 2001. He joined AT&T Labs - Research, Middletown, NJ, in 2000, and is currently a Principle Member of Technical Staff in the Video and Multimedia Technologies Research Department. He is an adjunct professor of the Electrical Engineering Department of Columbia University. His research interests include multimedia content analysis, multimedia databases, video search, pattern recognition, machine learning, and natural language understanding. He holds 28 U.S. patents and has published more than 60 papers in international conferences and journals. Dr. Liu is on the editorial board of the IEEE Transaction on Multimedia and the Peer-to-peer Networking and Applications Journal. He was also on the organizing committee and technical committee for a number of IEEE Conferences. Dr. Liu is a senior member of IEEE and a member of ACM.

Multimodal Video Copy Detection using Multi-Detectors Fusion

Yonghong Tian, Tiejun Huang, and Wen Gao (Fellow, IEEE)
Peking University, China
{yhtian, tjhuang, wgao}@pku.edu.cn

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.

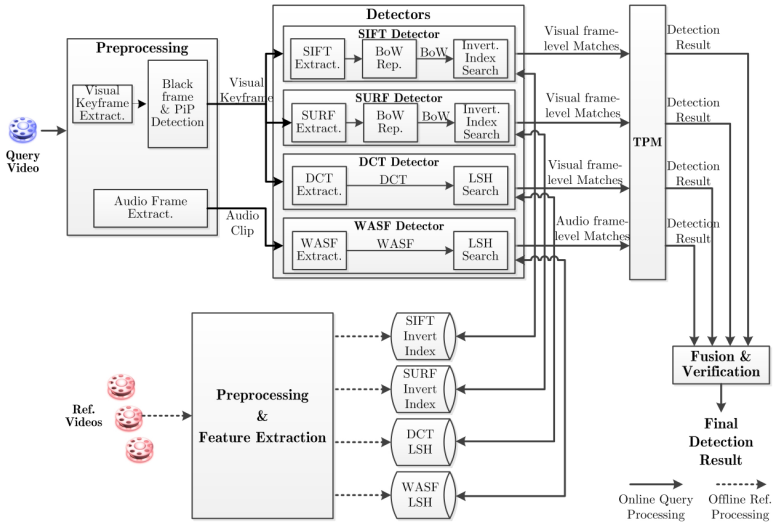


Figure 1. Framework of our copy detection system used for 2010 TRECVID-CBCD task

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) [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 [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, $\mathrm{D}_N = \langle d_1, d_2, \cdots, d_N \rangle$, 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 $s_1^{(\mathrm{V})}$ greater than or equal to a predefined threshold θ_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.

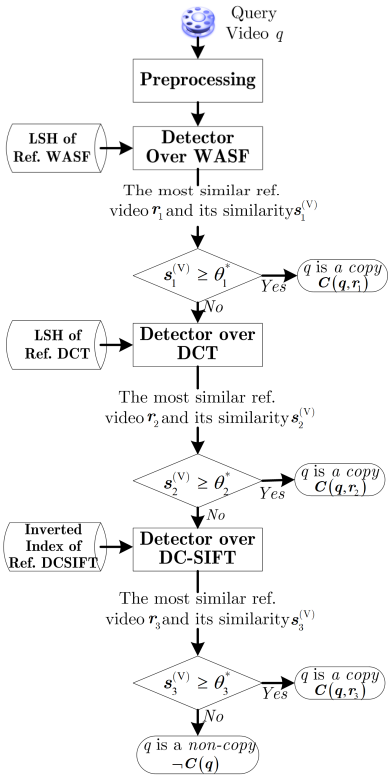


Figure 2. Schematic illustration of our CBCD-oriented cascading method used for the 2011 TRECVID-CBCD task

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.

	A1	A2	A3	A4	A5	A6	A7
V1	Case1: WASF Only				Case3: WASF+ DCT+DC-SIFT		
V2							
V3					Case2: WASF+DCT		
V4							
V5							
V6							
V8					Case3: WASF+ DCT+DC-SIFT		
V10							

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.

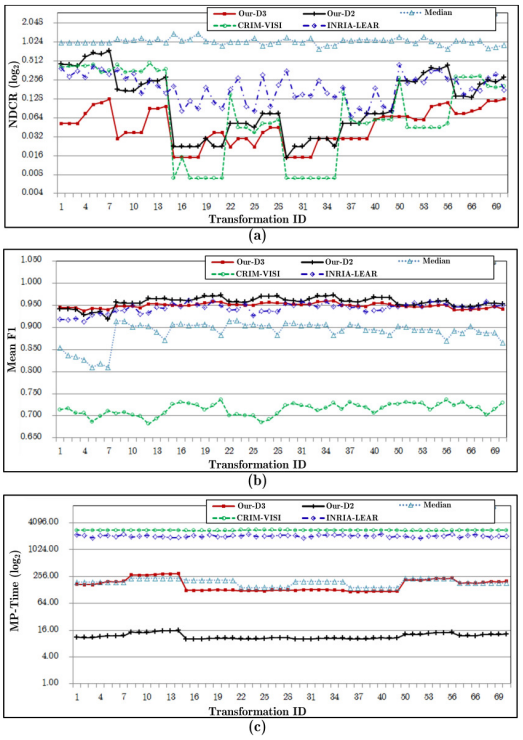


Figure 3. Performance curves of different methods over 56 transformations in the TRECVID-CBCD 2011 contest (for BALANCED profile only). (a) NDCR (y-axis in log2 coordinate), (b) Mean F1, (c) Mean Processing Time (y-axis in log coordinate).

However, it is obvious that artificial adjustment of decision thresholds for detectors can hardly achieve the optimal performance, and more importantly, lacks in generalization and is very burdensome. To fix this problem, a soft threshold learning algorithm was proposed in [10] to estimate the optimal decision thresholds for detectors. Experimental results showed that by utilizing soft threshold learning algorithm rather than manually tuning the thresholds, the effectiveness and efficiency can be further enhanced.

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

[1] T.J. Huang, Y.H. Tian, W. Gao, and J. Lu, Mediaprinting: identifying multimedia content for digital rights management, *Computer*, 43(12), 2010, 28-35.

[2] W. Kraaij, and G. Awad, TRECVID 2011 Content-based copy detection: task overview, Nov, 2011. http://www-nlpir.nist.gov/projects/tvpubs/tv11.slides/ tv11.ccd.slides.pdf.

[3] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *Int'l J. Computer Vision*, 60(2), 2004, 91-110.

[4] H. Bay, T. Tuytelaars, and L. V. Gool, SURF: Speeded Up Robust Features, *Proc. ECCV'06*, Vol. 3951, 2006, 404-417.

[5] J.P. Chen and T.J. Huang, A robust feature extraction algorithm for audio fingerprinting, *Proc. PCM'08*, 2008, 887-890.

[6] A. Gionis, P. Indyk, and R. Motwani, Similarity Search in High Dimensions via Hashing, *Prof. VLDB'99*, 1999, 518-529.

[7] Y.H. Tian, M.L. Jiang, L.T. Mou, X.Y. Fang, and T.J. Huang, A multimodal video copy detection approach with sequential pyramid matching, *Proc. ICIP'11*, 2011, 3629-3632.

[8] W. Kraaij, and G. Awad, TRECVID 2010 Content-based copy detection: task overview, Dec. 2010. http://www-nlpir.nist.gov/projects/tvpubs/tv10.slides/ tv10.ccd.slides.pdf.

[9] A. Bosch, A. Zisserman, and X. Muoz, Scene classification using a hybrid generative/discriminative approach, *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(4), 2008, 712-727.

[10] M.L. Jiang, Y.H. Tian, and T.J. Huang, Video Copy Detection Using a Soft Cascade of Multimodal Features, *Prof. ICME'12*, 2012, 374-379.

[11] C.-Y. Chiu, H.-M. Wang, and C.-S. Chen, Fast Min-Hashing Indexing and Robust Spatio-Temporal Matching for Detecting Video Copies, *ACM Trans. Multimedia Comput. Commun. Appl.*, 6(2) Article 10, March 2010, 23 pages.

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.

Near-Duplicate Image Retrieval as a Classification Problem

Yusuke Uchida and Shigeuki Sakazawa
KDDI R&D Laboratories, Inc., Japan
{ys-uchida, sakazawa}@kddilabs.jp

1. Introduction

With recent advances in both stable interest region detectors [1] and robust and distinctive descriptors [2], the retrieval of near-duplicate images has attracted significant attention [3]. It has become applicable to large-scale databases owing to the bag-of-visual words (BoVW) framework [4]. In the BoVW framework, local feature points or regions are first detected in an image, and then feature descriptors are extracted from them. These feature vectors are quantized into visual words (VWs) using a visual codebook, resulting in a histogram representation of VWs. In many cases, image similarity is measured by the L_2 distance between the normalized histograms. As the histograms are generally sparse, an inverted index data structure and a voting function enable an efficient similarity search. The equivalency between L_2 distances and scores obtained with the voting function is described in [5] in detail. As a way to emphasize distinctive VWs, the inverse document frequency (IDF) scoring [4] has been widely used, and shown to be effective.

Though the BoVW framework realizes efficient retrieval, some degradation of accuracy is caused by quantization [6]. Two major approaches are proposed to alleviate quantization error: post-filtering [5,8] and multiple assignment [5,9,10]. In post-filtering, after the VW-based matching any unreliable matches are filtered out according to the estimated distances between query and reference descriptors. In the multiple assignment approaches, a query descriptor can be matched not only with reference descriptors in the nearest VW, but also with reference descriptors in the several nearest VWs. Although quantization error is alleviated, all the above methods still depend on IDF scoring in voting, which is designed for words or quantized descriptors. In other words, the score is still *quantized*.

In this paper, we present a new scoring method applicable to both quantized and unquantized descriptors. The proposed score is based on the ratio of the probability density functions of an object model and a background model, which is efficiently calculated in an on-the-fly manner via nearest-neighbor density estimation. In experiments, we show the effectiveness of the proposed scoring method by applying it to a BoVW framework and a *k*-nearest neighbor voting framework.

2. Image retrieval as a classification problem

In this section, we first present the formulation of the proposed scoring method, starting with a classification problem. Then, in order to make it applicable to large-scale image retrieval, an approximation is introduced. Finally, the score is calculated via non-parametric density ratio estimation.

Probabilistic formulation.

Given a query image Q, the objective is to find a similar image $R_{j'}$ among a large number of reference images R_1, ..., R_C. Considering this as a classification problem, we start with maximum-a-posteriori estimation: $j' = \mathrm{argmax}_j\ p(R_j|Q)$. Assuming $p(R_j)$ is uniform, the maximum-a-posteriori estimation reduces to a maximum likelihood estimation:

$$j' = \arg\max_j p(\mathrm{R}_j \mid \mathrm{Q}) = \arg\max_j p(\mathrm{Q} \mid \mathrm{R}_j).$$

Letting $Q = \{q_1, ..., q_n\}$ denote the descriptors of the query image Q, with the naive Bayes assumption we get:

$$p(\mathrm{Q} \mid \mathrm{R}_j) = p(q_1, \cdots, q_n \mid \mathrm{R}_j) = \prod_{i=1}^{n} p(q_i \mid \mathrm{R}_j).$$

As pointed out in [6], if we assume all query descriptors are derived from only the object model of R_j, $p(Q|R_j)$ tends to be too small even if Q and R_j share the same object. In [6], the problem is alleviated by estimating $p(q_i|R_j)$ using a few dozen images representing the same class. As this is not practical for large-scale image retrieval, we model $p(q_i|R_j)$ with a mixture of the object model of R_j and a background model distinct from R_j:

$$p(q_i \mid \mathrm{R}_j) = \lambda p(q_i \mid R_j) + (1-\lambda) p(q_i),$$

where λ ($0 < \lambda < 1$) is the mixture parameter, and R_j denotes a set of descriptors in the reference image R_j. If we consider the descriptors Q and R_j as words, this is identical to LM (language modeling)-RSV [11] in the area of information retrieval (IR). Combining the above equations, we obtain:

$$\begin{aligned} j' &= \arg\max_j \prod_{i=1}^{n} p(q_i \mid \mathrm{R}_j) \\ &= \arg\max_j \sum_{i=1}^{n} \log(\lambda p(q_i \mid R_j) + (1-\lambda) p(q_i)) \\ &= \arg\max_j \sum_{i=1}^{n} \log\left(\frac{\lambda}{1-\lambda} \frac{p(q_i \mid R_j)}{p(q_i)} + 1\right). \end{aligned}$$

Representing $\log\left(\frac{\lambda}{1-\lambda} \frac{p(q_i \mid R_j)}{p(q_i)} + 1\right)$ by s_{ij}, the objective is now to identify the reference image R_j with the largest $\Sigma_i S_{ij}$, where s_{ij} can be considered to be a voting score reflecting the contribution of the descriptor q_i to the image R_j.

Approximation with nearest neighbors.
The above formulation requires the calculation of s_{ij} for all R_j. Letting C denote the number of classes and D denote the average number of descriptors in an image, the calculation of score s_{ij} for all R_j has a time cost of $O(C\log(D))$ with efficient (approximate) nearest-neighbor search algorithms [8,12,13]. This does not become a fatal flaw in classification problems where $D>>C$. However, it is intractable in large-scale image retrieval problems where $C>>D$, where C corresponds to the number of images or objects in a database. In order to make it tractable, the following simple approximation is adopted. We assume the nearest-neighbor descriptors $N(q_i)$ of q_i (e.g., k-nearest neighbors of q_i) were obtained against all reference descriptors. Then, $p(q_i|R_j)$ is calculated *only* for R_j, at least one of whose descriptors appears in $N(q_i)$, and otherwise we assume $p(q_i|R_j) = 0$. Because s_{ij} becomes 0 if $p(q_i|R_j) = 0$, the calculation of $\Sigma_i S_{ij}$ is performed efficiently. With this approximation, the computational cost is reduced from $O(C\log(D))$ to $O(\log(CD))$.

Non-parametric density ratio estimation.
Finally, s_{ij} is calculated using $N(q_i)$. We assume that $N(q_i)$ has t subsets

$$N_1(q_i) \subset N_2(q_i) \subset \cdots \subset N_t(q_i) = N(q_i),$$

such that they satisfy

$$a < b \quad \Rightarrow \quad p(q_i \mid N_a(q_i)) > p(q_i \mid N_b(q_i)).$$

For each $N_s(q_i)$ ($1 \leqq s \leqq t$), the densities $p(q_i|R_j)$ and $p(q_i)$ are estimated via k-nearest neighbor density estimation only for R_j, one of whose descriptors appears in $N_s(q_i)$:

$$p(q_i \mid R_j) = \frac{|N_s(q_i)|_j}{|R_j| \cdot V_s}, \quad p(q_i) = \frac{|N_s(q_i)|_{\text{all}}}{|R_{\text{all}}| \cdot V_s},$$

where $|N_s(q_i)|_j$ is the number of descriptors of R_j that appear in $N_s(q_i)$, R_{all} is the set of all reference descriptors, V_s is the volume of a hyper-sphere with radius $|q_i - r'_s|$, and r'_s is the farthest descriptor from q_i in $N_s(q_i)$. Finally, we obtain:

$$s_{ij} = \log\left(\frac{\lambda}{1-\lambda} \frac{|N_s(q_i)|_j / |R_j|}{|N_s(q_i)|_{\text{all}} / |R_{\text{all}}|} + 1\right).$$

Although a score s_{ij} can be obtained for each s, we adopt $\max_s s_{ij}$ as the final score. More concrete examples of the formulation are shown in the following section.

3. Performance evaluation
In this section, we show the effectiveness of the classification-based method by applying it to the BoVW and k-nearest neighbor (k-NN) voting frameworks. Experiments were performed on the University of Kentucky recognition benchmark dataset [14], which includes 2,550 different objects or scenes. Each of these objects is represented by four images taken from four different angles, making 10,200 images. These images are used as both reference and query images. Mean average precision (MAP) [5,14] is used as an indicator of retrieval performance. The detailed parameters are shown in [5,7,15].

Application to the BoVW-based framework.
First, we show the result when the above scheme is applied to the traditional BoVW framework. In this case, the nearest-neighbor descriptors $N_s(q_i)$ of q_i are defined as the descriptors that are assigned to the s nearest VWs of q_i, and m corresponds to the number of multiple assignments in multiple assignment approaches [5,9,10]. In the case where $m = 1$, this becomes the same as the BoVW framework except for the scoring method. Figure 1 shows the MAP scores obtained with the proposed method (PROP) and the traditional IDF-based scoring method (IDF) as a function of λ. We can see that the proposed scoring method significantly improves the performance. The best MAP score of 0.814 is achieved with relatively low λ ($\lambda = 0.07$), which implies that there are a small number of features useful for object recognition [16].

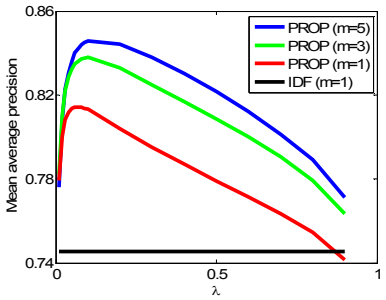


Figure 1. Performance of BoVW-based image retrieval.

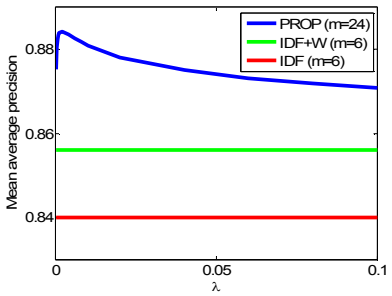


Figure 2. Performance of k-NN-based image retrieval.

Application to the k-NN-based framework.
Next, we apply the proposed scheme to the k-NN voting framework. In this case, the nearest neighbor descriptors $N_s(q_i)$ are defined simply as the s nearest descriptors of q_i, and m is the number of the nearest neighbors used in the voting. In Figure 2, we also compare the method (IDF+W) which utilizes the distance d between the descriptors as a weight (e.g.,

$\exp(-d^2/\sigma^2)$) in the voting [5,7]. In Figure 2, it is also shown that the proposed scoring method improves the accuracy. In this experiment, *m* has been optimized for each method independently.

4. Conclusion

In this paper we have proposed a new scoring method for local feature-based image retrieval, which is based on the ratio of the probability density function of an object model to that of a background model. The effectiveness of the proposed method was confirmed by applying it to the bag-of-visual words-based framework and the *k*-NN voting framework. The proposed approach can be applied to many problems where content is represented by a set of descriptors, and the neighbors of the descriptors can be obtained efficiently under any metrics.

References

[1] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," in *IJCV*, vol. 60, no. 1-2, pp. 43-72, 2005.

[2] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," in *TPAMI*, vol. 27, no. 10, pp. 1615-1630, 2005.

[3] H. Xie, K. Gao, Y. Zhang, S. Tang, J. Li, and Y. Liu, "Efficient Feature Detection and Effective Post-Verification for Large Scale Near-Duplicate Image Search," in *TMM*, vol. 13, no. 6, pp. 1319-1332, 2011.

[4] J. Sivic and A. Zissermane, "Video google: A text retrieval approach to object matching in videos," in *Proc. of ICCV*, pp. 1470-1477, 2003.

[5] H. Jegou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," in *IJCV*, vol. 87, no. 3, pp. 316-336, 2010.

[6] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. of CVPR*, pp. 1-8, 2008.

[7] Y. Uchida, K. Takagi, and S. Sakazawa, "Ratio Voting: A New Voting Strategy for Large-Scale Image Retrieval," in *Proc. of ICME*, 2012.

[8] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," in *TPAMI*, vol. 33, no. 1, pp. 117-128, 2011.

[9] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. of CVPR*, pp. 1-8, 2008.

[10] A. Mikulik, M. Perdoch, O. Chum, and J. Matas, "Learning a fine vocabulary," in *Proc. of ECCV*, pp. 1-14, 2010.

[11] T. Roelleke and J. Wang, "Tf-idf uncovered: A study of theories and probabilities," in *Proc. of SIGIR*, pp. 435-442, 2008.

[12] A. Andoni, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," in *Proc. of FOCS*, pp. 459-468, 2006.

[13] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *Proc. of VISAPP*, pp. 331-340, 2009.

[14] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. of CVPR*, pp. 2161-2168, 2006.

[15] Y. Uchida, K. Takagi, and S. Sakazawa, "An Alternative to IDF: Effective Scoring for Accurate Image Retrieval with Non-Parametric Density Ratio Estimation," in *Proc. of ICPR*, 2012.

[16] P. Turcot and D. G. Lowe, "Better matching with fewer features: The selection of useful features," in *Proc. of WS-LAVD*, 2009.

Yusuke Uchida received a Bachelor's Degree in Integrated Human Studies from Kyoto University, Kyoto, Japan, in 2005. He received the degree of Master of Informatics from the Graduate School of Informatics, Kyoto University, in 2007. His research interests include large-scale content-based multimedia retrieval, augmented reality, and image processing. He is currently with KDDI R&D Laboratories, Inc.

Shigeyuki Sakazawa received his B.E., M.E., and Ph.D. degrees from Kobe University, Japan, all in electrical engineering, in 1990, 1992, and 2005 respectively. He joined Kokusai Denshin Denwa (KDD) Co. Ltd. in 1992. Since then he has been with its R&D Division, and now he is a senior manager of the Media and HTML5 Application Laboratory of KDDI R&D Laboratories Inc., Saitama, Japan. His current research interests include video coding, video communication systems, image recognition and CG video generation.

Robust Media Search Technology for Content-Based Audio and Video Identification

Ryo Mukai, Takayuki Kurozumi, Takahito Kawanishi, Hidehisa Nagano and Kunio Kashino
NTT Corporation, Japan
{mukai.ryo, kurozumi.takayuki, kawanishi.takahito, nagano.hidehisa, kashino.kunio}@lab.ntt.co.jp

1. Introduction

The increase in the capacity of networks and storage devices is causing an explosive expansion in the volume of media data that we can access. Accordingly, demand is rapidly growing for tools that allow us to search for and identify media information using not only key words but also audio, video or image content. We consider media search to be a core technology for meeting such needs. The most basic function of a media search involves detecting and locating media fragments that are similar to a media fragment presented as a query in a huge number of media archives. This function is sometimes called content-based copy detection. The copy detection of media data has a wide range of applications including copyright management and enforcement, derivative tracking, advertising and recommendations.

In media search, robustness is particularly important because media signals are often converted to various encoding formats, mixed with other signals such as background music, or even edited and re-edited into different versions. Search speed is also crucial, considering the rapidly growing volumes of audio and video being created, distributed and exchanged by individuals, public institutions and corporations around the world.

2. Robust Media Search (RMS) Technology

Robust Media Search (RMS) is a core technology for content-based audio and video media search and identification developed by NTT. It offers excellent robustness and a very high search speed by using coarsely-quantized features and spatiotemporal consistency. As shown in Figure 1, the media data are converted to sequences of coarsely-quantized digits called “features” or “fingerprints”. Note that not all of these digits are necessarily used for matching; we found that appropriately choosing digits to be matched from among all the digits not only simplifies the search but also greatly improves the robustness of the search, which is performed by matching those features. Specifically, the spatiotemporal accumulation of many such features enables us to achieve extremely high identification accuracy.

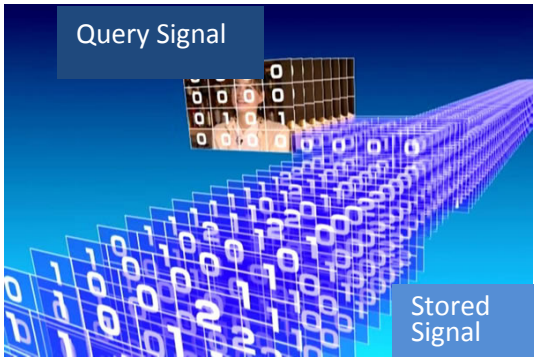


Figure 1. Image of video feature (fingerprint) of Robust Media Search

3. Algorithms

This section briefly reviews the algorithms used in RMS.

3.1 Divide-And-Locate method

The Divide-And-Locate method (DAL) is an audio fingerprinting technology, which is especially robust as regards additive noise [1]. The basic idea of the DAL is to divide a spectrogram into a number of small regions and undertake matching for each region to locate it in the database. To extract the feature, the spectrogram is divided into small regions (Figure 2). The spectrum corresponding to each region is classified by vector quantization (VQ). The feature extracted from each region then undergoes the matching process with the features of the signals in the database to locate a part that matches many regions. The matching operation is executed by looking up a similarity table among the VQ codes and scanning index lists made in advance from feature data of the reference signals. This operation is executed much more efficiently than an exhaustive search, therefore the DAL realizes a very fast search over a huge database. The detailed algorithm is described in [1, 2, 3].

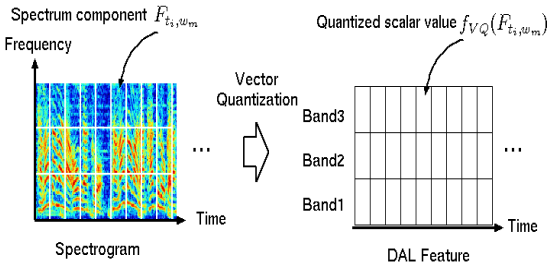


Figure 2. Feature extraction for Divide-And-Locate method

3.2 Binary Area Matching method

The Binary Area Matching method (BAM) was developed to cope with multiplicative noise or distortion. The BAM method first divides the signal into small segments, and then selects the portions in which the signal changes greatly over time. In other words, it selects only the segments that are rather unsusceptible to real environmental influences. Next, the selected segments are quantized coarsely and represented by binary numbers (Figure 3). By comparing the signal after coarse rather than fine quantization, the influence of distortion is weakened, thus achieving a robust search. The algorithm and experimental results are detailed in [4].

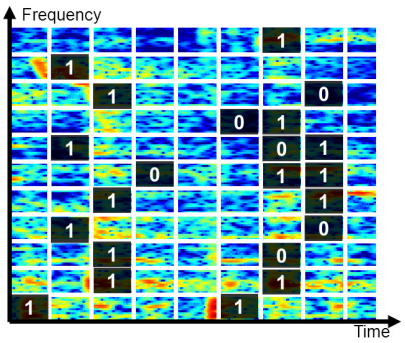


Figure 3. Feature data of Binary Area Matching

3.3 Coarsely-quantized Area Matching method

The Coarsely-quantized Area Matching method (CAM) is our video fingerprinting technology, which offers excellent robustness against various kinds of distortion. The algorithm is detailed in [2, 3, 5]. The CAM uses the color values of each pixel in the video stream. Each frame of the video is divided into small regions, and then the areas in which color changes greatly over time are selected. The feature of the CAM is the time series of the coarsely quantized pixel values of the selected areas.

Figure 4 shows the feature extraction procedure. First, the input video frame is divided into small regions. We adopt the average RGB values over each region as a primitive video feature. On the other hand, we calculate an average and a standard deviation of the pixel values over a time window. We assume an area that deviates greatly from a temporal local average to be a salient part. Based on this assumption, we make a mask for feature selection based on the standard deviations. Next, we quantize the selected feature values. The quantization is carried out on locally normalized feature values. Accordingly, we obtain the feature data of the CAM.

In the search stage, we perform a time-series search for the query feature in the reference feature database using codes obtained by quantizing locally normalized feature values. The database is scanned with a sliding window that has the same length as a query segment. Similarity is measured in terms of the Hamming distance between masked query codes and reference codes, that is, by counting the number of co-occurrences of the quantized codes at the corresponding positions in the window. This procedure is implemented very efficiently by using a hash table whose key is a pair consisting of the code and coordinates.

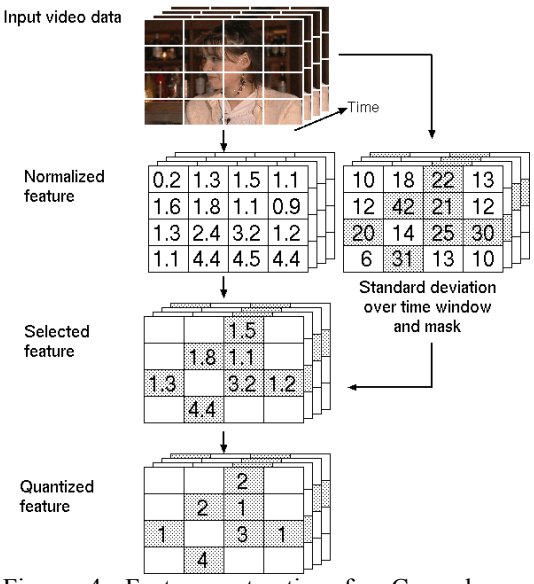


Figure 4. Feature extraction for Coarsely-quantized Area Matching

4. Applications

The RMS has a wide range of applications, and it has been already deployed for many commercial services.

Internet copyright enforcement

The RMS has been used in copyright monitoring systems for video sharing sites on the Internet [6]. The system crawls over 250k files/day posted on major video sharing sites, and it detects the use of known (previously fingerprinted) audio and/or video content in real time.

Media search with smartphones

Figure 5 shows a prototype system for video search on a smartphone. It captures a few seconds of video and identifies its title. This technology is applicable to various kinds of content navigation systems including content based information retrieval using smartphones

and advertisements synchronized with media data such as broadcast video, music, or commercial messages.

Figure 5. Video search on smartphone

Media Link Analysis

Media Link Analysis is a technique for automatically adding annotations to the vast quantity of media data. For example, for simultaneously recorded multiple-channel TV broadcasts, the analysis compares the latest video data with stored data. If any segments match, a link is created. The number of links means the number of times the segment has been used. The analysis therefore shows the usage count for each segment, and the counts can be used as an index of popularity or importance.

5. Conclusions

Media search will be a vital future information retrieval method. We expect that media search will be used for the media cloud [7], in which huge amounts of media data are stored, created, distributed, and consumed. Media search will be a core technology that explores the links between one form of media data and another, or media data and various kinds of information in the media cloud, based on the relationships between the parts of the media content, such as partial similarity or partial quotations. We anticipate that media search will be an essential technology if we are to fully utilize all the media information now exploding worldwide.

References

[1] H. Nagano, K. Kashino and H. Murase, "A Fast Search Algorithm for Background Music Signals Based on the Search for Numerous Small Signal Components," in *Proc. ICASSP 2003*.

[2] R. Mukai, T. Kurozumi, K. Hiramatsu, T. Kawanishi, H. Nagano and K. Kashino, "NTT Communication Science Laboratories at TRECVID 2010 Content-Based Copy Detection," in *Proc. TRECVID 2010*, http://www-nlpir.nist.gov/projects/tvpubs/tv10.papers/ntt-csl.pdf

[3] R. Mukai, T. Kurozumi, T. Kawanishi, H. Nagano and K. Kashino, "NTT Communication Science Laboratories at TRECVID 2011 Content-Based Copy Detection," in *Proc TRECVID 2011*, http://www-nlpir.nist.gov/projects/tvpubs/tv11.papers/ntt-csl.pdf

[4] K. Kashino, A. Kimura, H. Nagano and T. Kurozumi, "Robust Search Methods for Music Signals Based on Simple Representation," in *Proc. ICASSP 2007*.

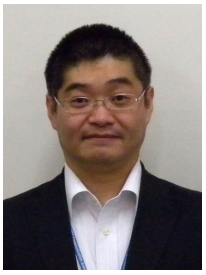
[5] T. Kurozumi, H. Nagano and K. Kashino, "A Robust Video Search Method for Video Signal Queries Captured in the Real World," *IEICE Trans. Inf.& Syst.(Japanese Edition)*, vol. J90-D, no. 8, pp. 2223-2231, 2007, in Japanese.

[6] NTT Data Corp. Press Release, "New Content Monitoring Service Identifies Audio and Video on the Internet Quickly and Accurately," 2009, http://www.nttdata.com/global/en/news-center/global/2008/120100.html

[7] K. Kashino, "Robust Media Search in the Cloud," in *Proc. GJS 2010*, http://www.gjs2010.org/docs/S3-3_K.Kashino.pdf

Ryo Mukai received B.S. and M.S. degrees in information science from the University of Tokyo, Japan, in 1990 and 1992, respectively. He joined NTT Corporation in 1992. From 1992 to 2000, he was engaged in the research and development of processor architecture for network service systems and distributed network systems. Since 2000, he has been with NTT Communication Science Laboratories, where he is engaged in research on media search technology. He is a senior member of the IEEE, and a member of the ACM, ASJ, IEICE and IPSJ.

Takayuki Kurozumi received a B.S. degree in physics from the Tokyo Metropolitan University, Japan, in 1997, and M.S. and Ph.D. degrees in information science from the Japan Advanced Institute of Science and Technology, Ishikawa, in 1999 and 2007, respectively. In 1999, he joined Nippon Telegraph and Telephone Corporation, Atsugi, Japan. His research interests include pattern recognition, image processing, and multimedia information retrieval. He is a senior member of the IEICE.



Takahito Kawanishi is Senior Research Scientist, Research Planning Section at NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation. He received a B.E. degree in information science from Kyoto University, Kyoto, M.E and Ph.D. degrees in information science from Nara Institute of Science and Technology, Nara, in 1996, 1998, and 2006, respectively. He joined NTT Laboratories in 1998. From 2004 to 2008, he worked in Plala Networks Inc. (Now NTT Plala) as a technical manager and developed commercial IPTV and VoD systems. He is currently engaged in R&D of online media content identification, monitoring and search systems. He is a senior member of IEICE, and a member of IPSJ and JSIAM.

Hidehisa Nagano received B.Eng. and M.Eng. degrees in information and computer sciences in 1994 and 1996, respectively, and a Ph.D. degree in information science and technology in 2005, all from Osaka University. In 1996, he joined NTT, where he is currently Senior Research Scientist of the Media Information Laboratory, NTT Communication Science Laboratories. From 2011 to 2012, he was a visiting researcher at the Centre for Digital Music, Queen Mary University of London, UK. He has been working on audio and video analysis, search, retrieval, and recognition algorithms and their implementation. He is a senior member of the IEEE, and a member of the IEICE and IPSJ.

Kunio Kashino is the Leader of the Media Recognition Research Group at NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation. He has been working on audio and video analysis, search, retrieval, and recognition algorithms and their implementation. He received a PhD from the University of Tokyo for his work on music scene analysis in 1995. He is a Senior Member of the IEEE.

Multimodal Video Copy Detection using Local Features

Xavier Anguera[1] *and Tomasz Adamek*[2]
[1]*Telefónica Research, Spain*
[2]*Catchoom, Spain*
xanguera@tid.es

1. Introduction

Content-based video copy detection (CBCD) systems aim at finding video segments that are identical or transformed versions of segments in a known video. Joly et al. [8] propose a definition of video copy based on a subjective notion of tolerated transformations. A tolerated transformation is a function that creates a new version of a document where the original document "remains recognizable". CBCD systems perform the detection by processing visual and/or audio content of videos, ignoring any metadata and avoiding the embedding of watermarks into the original videos.

The detection of video duplicates in a video database is one of the key technologies in multimedia management. Its main applications include, among others: (a) storage optimization; (b) copyright enforcement; (c) improved web search and (d) concept tracking. In storage optimization, the goal is to eliminate exact duplicate videos from a database and to replace them with links to a single copy or, alternatively, link together similar videos (near duplicate) for fast retrieval and enhanced navigation. Copyright enforcement strives at avoiding the use and sharing of illegal copies of a copyright protected video. In the context of web search, the goal is to increase novelty in the video search result list, by eliminating copies of the same video that may clutter the results and hinder users from finding the desired content [9]. Finally, concept tracking in video feeds [10] focuses on finding the relationship between segments in several video feeds, in order to understand the temporal evolution of, for example, news stories.

Traditionally, CBCD systems have relied only on the visual information in the videos, driving research towards new features that allowed for scalable systems and fast retrieval. In recent years systems started incorporating features derived from the audio modality. Through the fusion of multimodal information, CBCD systems can obtain improved performance levels and be more robust to errors (e.g. when one of the modalities is severely damaged or missing).

In this letter we briefly describe the CBCD system we have developed over the last 3 years at Telefónica Research, which has achieved outstanding results in the 2010 and 2011 NIST-Trecvid Video Copy Detection evaluations. The main characteristics of our system are the use of multimodal (i.e. audio and video) information to detect plausible video copies in each modality and an effective late fusion of results.

The Individual modalities we use are both based on local features (DART [4] for video and MASK [6] for audio) which are able to robustly encode the multimodal content, thus allowing for successful retrieval of video copies. The late fusion algorithm [7] takes into account the scores and ranking of each result in each modality to combine them most effectively. The currently system has been developed into a pseudo-commercial application by using a scalable database architecture and feature extraction parameters optimized for speed.

2. Multimodal Video Copy Detection System

Figure 1 shows the main blocks that conform Telefónica Research CBCD system. The system is composed of two parallel feature streams (one for audio and one for video) that are processed in parallel, obtaining each one an individual set of results (for NIST-Trecvid experiments we retrieve 20 results from each modality). Then a late fusion is used to merge these results into a multimodal output. The choice of a late fusion versus an earlier one was done so that the system gains in flexibility to be applied for different applications like music IR, where one of the modalities is non-existent or relevant.

Next we describe each of the steps in the system.

2.1 Local Video System

The local visual features processing module [1][5] compares all query keyframes with all reference keyframes using a state-of-the-art image retrieval engine relying on local features [4] and then combines the obtained ranked lists of matched keyframes into copied video segments by performing a temporal consistency post-processing. The module is divided in three tasks: Feature Extraction (which samples every video with one frame per second and extracts novel local features called DART [4]), Keyframe Matching (by using hierarchical dictionaries of visual words and inverted files to locate matching frames, which are then refined through a spatial verification stage), and Temporal Consistency (which computes the time differences between matches, and returns a list copy

candidates each one with a location within the video, and a score).

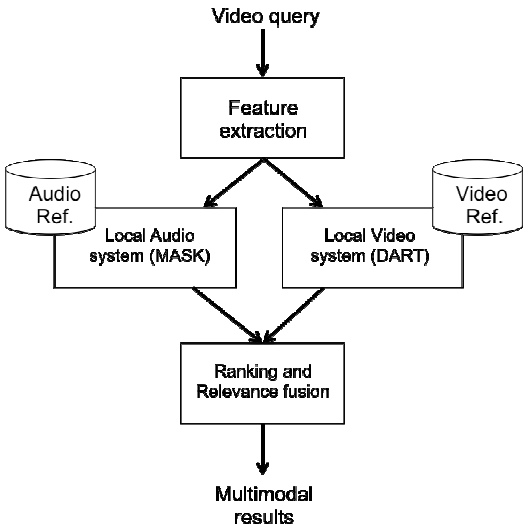


Figure 1: Multimodal Video Copy Detection System

2.2 Local Audio System

The audio local features module [5] is based on the recently proposed MASK features [6]. These are local features computed over the spectral domain of the audio signal by encoding in binary form the energy differences of predefined regions around spectral maxima. The MASK features therefore consist on an asynchronous stream of fingerprints that later are matched, using a similar technique to the one used in the video system, with the reference features. Instead of a spatial consistency step, in here we perform an alignment step between all MASK points in the reference and query matching segments to obtain an accurate matching score.

2.3 Ranking and relevance fusion

Each of the modalities in the presented system results in a list of 20-best possible matches between segments in the provided query video and the reference database. The fusion module is in charge of merging these monomodal decisions into a multimodal output as described in [7]. To do so, it makes use of both the ranking of each matching segment within its modality and the normalized score. As each modality might have different score distributions and sometimes not even output all 20 matches, an L1 normalization and flooring preprocessing is performed before the results are merged.

In [5] we show the suitability of this fusion algorithm for any number of monomodal inputs. We show we are able to lower the best score obtained in the NIST-Trecvid evaluation by a blind combination of several of the submitted system outputs without.

2.4 Scalable implementation

Ultimately, a CBCD system is useful if it can be scaled to index large multimodal databases and it is able to retrieve matching segments for a given query in little time. In order to achieve the first goal we use a disk-based inverted file index over SSD drives that structures the information at indexing time in a way that it is faster to retrieve similar reference matches at retrieval time. Also, given the locality of both fingerprints we use, we are able to adapt the amount of information we store per second of content depending on the application and the relation between accuracy and speed we want to achieve.

Currently our system is flexible to accommodate various application use-cases as neither we impose constraints on the length of the query or reference, nor we predefine where the matching start-end points might be found.

References

[1] E. Younessian, X. Anguera, T. Adamek, N. Oliver, and D. Marimon, “Telefonica research at trecvid 2010 content-based copy detection,” in Proc. NIST-TRECVID Workshop, 2010.

[2] T. Adamek and D. Marimon, “Large-scale visual search based on voting in reduced pose space with application to mobile search and video collections,” in Multimedia and Expo (ICME), 2011 IEEE International Conference on, july 2011, pp. 1 –4.

[3] X. Anguera, J. M. Barrios, T. Adamek, and N. Oliver, “Multimodal fusion for video copy detection,” in Proc. ACM Multimedia, 2011.

[4] D. Marimon, A. Bonnin, T. Adamek and R. Gimeno, “DARTs: Efficient scale-space extraction of daisy keypoints,” in Proc. Computer Vision and Pattern Recognition (CVPR), June 2009.

[5] X. Anguera, T. Adamek, D. Xu and J.-M. Barrios, “Telefonica Research at TRECVID 2011 Content-Based Copy Detection”, in Proc. NIST-TRECVID Workshop, 2011

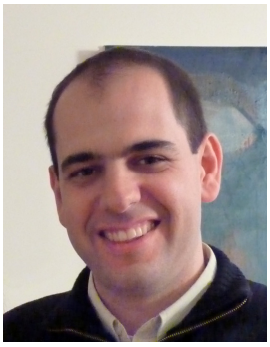
[6] X. Anguera, A Garzon and T. Adamek, “MASK: Robust Local Features for Audio Fingerprinting”, in Proc. ICME 2012, Melbourne, Australia

[7] X. Anguera and J-M. Barrios, T. Adamek and N. Oliver, “Multimodal Fusion for Video Copy Detection”, in Proc. ACM Multimedia 2011.

[8] A. Joly, O. Buisson, and C. Fr´elicot. Content-based copy retrieval using distortion-based probabilistic similarity search. IEEE Trans. on Multimedia, 9(2):293–306, 2007.

[9] X. Wu and A. G. Hauptmann and Ch.-W. Ngo, “Novelty Detection for Cross-Lingual News Stories with Visual Duplicates and Speech Transcripts”, in Proc. ACM Multimedia 2007

[10] Y. Zhai and M. Shah, "Tracking News Stories Across Different Sources", In Proc. ACM Multimedia 2005.



Xavier Anguera Ing. [MS] 2001 by UPC (Barcelona, Spain), [MS] 2001 European Masters in Language and Speech, Dr. [PhD] 2006 UPC University. From 2001 to 2003 he worked for Panasonic Speech Technology Lab in Santa Barbara, CA. From 2004 to 2006 he was a visiting researcher at the International Computer Science Institute (ICSI) in Berkeley, CA. Since 2007 he is with Telefonica Research in Barcelona, pursuing research on multimedia analysis. His main research interests involve speech/speaker processing and multimedia processing.

Tomasz Adamek is the cofounder and CTO of Catchoom, a startup that provides visual recognition technology designed for real live applications. He received his Ph.D. degree from Dublin City University (DCU), Ireland, in 2006. Between 2006 and 2007 he worked as a postdoctoral researcher at CDVP, DCU. In 2008 he joined Telefónica R&D, Barcelona, Spain where he was responsible for technological aspects of several research projects in the area of large-scale multimedia indexing and retrieval. His work led to several patents and over 20 academic papers in high impact journals and conferences.

Datasets for Content-Based Video Copy Detection Research

Yudong Jiang and Yu-Gang Jiang
Fudan University, China
{11210240091, ygj}@fudan.edu.cn

1. Introduction

The explosive growth of the Internet video sharing activity has largely amplified the longstanding data copyright issue. *Content-Based Copy Detection* (CBCD) system aims to automatically identify video copies in large-scale databases, which has received significant research attention in multimedia and computer vision. Typically a CBCD system first extracts various visual and audio features and then builds an indexing structure over large databases for efficient copy search. Many algorithms and systems have been proposed to tackle this challenge. Readers are referred to the online notebook papers of NIST TRECVID evaluations [1] for a significant portion of the state-of-the-art techniques.

In this short paper we do not intend to discuss CBCD techniques. Instead we provide a compact overview on benchmark datasets in CBCD research. Like any multimedia and vision tasks, good benchmark dataset is one of the central factors that surely affect the advancement of techniques. In the following we discuss the pros and cons of current popular benchmarks, and briefly introduce a new benchmark that is current under construction in our group.

2. Current CBCD Datasets

Although video copy detection has been investigated for decades, before year 2007 there was basically no widely adopted benchmark. Normally people constructed their own dataset for research and did not release the data for cross-site comparison. For instance, Indyk et al. downloaded 2000 MPEG clips of news, music videos, and movie trailers [2]. The duration of these clips was typically between 2 and 5 minutes. Copies were generated by themselves, using "artificial" transformations including inserting TV logos and using various camcordings, frame rates, and different video formats. Although this dataset was pretty well-defined at the time of construction, it does not reflect the actual scenario in large-scale and complex circumstances as of today. Joly et al. collected 1,040 hours of TV video data stored in MPEG1 format, containing content in various categories like commercials, news, sports, and TV shows. Copies were also created artificially by applying several predefined transformations.

The datasets mentioned above have successfully served as the basis for algorithm evaluation in just a single paper/work. However, without a public benchmark for cross-site evaluation it is very difficult to push CBCD technologies forward. Perhaps the first well-known public benchmark is Muscle-VCD-2007 [4], created by Law-To et al. during ACM CIVR 2007 conference where a competition on CBCD was organized. The dataset contains around 100 hours of videos from the Internet, TV archives, and movies. Videos are in different resolutions and video formats. There are two kinds of queries representing two practical situations: (1) ST1: whole video copy (normally between 5 minutes to 1 hour duration), where the videos may be slightly recoded and/or noised. (2) ST2: partial video copy, where two videos only share one or more shot segments. This scenario was simulated by choosing shot segments from the dataset and using video-editing softwares to artificially create a few transformations. The "transformed" segments were later used as query to search the original version in the dataset. People not only should find the original videos where the segments were extracted, but also need to localize the segments with their start and end time codes. The duration of a segment normally ranges from 1 second to 1 minute.

Muscle-VCD-2007 is a ground-breaking effort in CBCD research since it provides probably the first open and well-defined benchmark, which has been evidenced by the recent progress of CBCD techniques, which were mostly evaluated on this dataset. Nevertheless, as the amount of online videos grows explosively, using only 100 hours of videos to test CBCD methods is apparently not sufficient to simulate the real-world copy detection challenge.

The importance of CBCD task was also recognized by the U.S. National Institute of Standards and Technology (NIST), whose annual TRECVID evaluation [1] created a separate CBCD task for the first time in 2008, to stimulate research in video copy detection. The evaluation was conducted yearly until 2011, with a benchmark dataset generated each year. The datasets are only available to registered participants of the task, which were constructed in a very similar way to Muscle-VCD-2007. The 2008 edition of the TRECVID CBCD dataset contains 200 hours of Dutch documentary videos, and around 2000

query clips. Each query was generated using software to randomly extract a segment from the dataset and add in some pre-defined transformations.

Additionally, the publicly available CC_Web _Video dataset constructed in 2007 by City University of Hong Kong and Carnegie Mellon University has also been widely used in the community. Although the dataset was created for near-duplicate video detection problem, which is slightly different from the copy detection task by definition, most of the *near-duplicates* are also copies. The near-duplicate videos were all from YouTube (not by artificial transformations), but are quite simple (mostly copies are duplicates of an entire video) when compared with the cases in Muscle-VCD-2007 and TRECVID datasets.

3. A More Realistic CBCD Dataset

As can be seen from the short discussion in Section 2, almost all the existing CBCD datasets were generated artificially using softwares to embed a number of video transformations. This partially satisfies the goals for evaluating new algorithms and systems, but also has some obvious drawbacks. First, the pre-defined transformations are still limited and cannot cover complex video copy cases from the real world. Second, some queries generated in this way from the TRECVID practice were transformed way too much. For example, some queries were seriously blurred and recaptured, which is very rare in reality since—after all—the video copies are intended to be viewed by people.

In view of the limitations of the current CBCD datasets, we are constructing a new benchmark, with which we hope to drive CBCD research into a new era. Videos in this new dataset were all downloaded from Internet video sharing sites YouTube and MetaCafe [6]. We do not generate artificial video copies using the pre-defined transformations. All copies in this dataset are *real data* directly from the Internet, which cover around 50 topics including news, sports, film, TV program, music video, cartoon, speech and so on. Some example copy frames in our dataset are shown in Figure 1.

Each topic has about 20 videos which are mostly copies of each other. Most of the copies are only at very short segment level, which makes the ground-truth annotation process extremely challenging and time-consuming. This is also one of the most important reasons that there's still no real video copy dataset in the community. During manual annotation, we take each video as a reference and mark all its (partial) copies in the other videos within the same topic. In total we need to identify partial copy segments between 10,000 video pairs, which will cost over 1000 human hours according to our estimation. We have designed an annotation tool to speed up this process, which utilized some heuristics such as the transitivity property.

Figure 1. Two example copy frame pairs in our new dataset.

After annotating the copies between the videos in the 50 topics, we will add in a large number of background videos to make this task more realistic and challenging. This dataset will be released upon its completion, expected sometime in early 2013.

As the performance of state-of-the-art methods tend to saturate on the current datasets, we hope that—with this new, realistic, and challenging benchmark—research in CBCD will gather some more momentum.

[*The release of this new dataset will be announced on www.yugangjiang.info. Stay tuned!*]

References

[1] http://trecvid.nist.gov/

[2] P. Indyk, G. Iyengar, and N. Shivakumar, "Finding pirated video sequences on the Internet." Technical report, Stanford University, 1999.

[3] A. Joly, C. Frélicot, and O. Buisson, "Robust content-based video copy identification in a large reference database," In Proceedings of International Conference on Image and Video Retrieval, 2003.

[4] http://www-rocq.inria.fr/imedia/civr-bench/

[5] http://vireo.cs.cityu.edu.hk/webvideo/

[6] http://www.metacafe.com/

Yudong Jiang received the Bachelor degree in Computer Science from Shenyang Aerospace University, Shenyang, China, in 2009. During 2009-2011, he worked in the software industry and developed two popular Social Games. He is currently pursuing the Master degree in Computer Engineering at Fudan University, Shanghai, China. His research interests include computer vision and machine learning.

Yu-Gang Jiang received the Ph.D. degree in Computer Science from the City University of Hong Kong, Kowloon, Hong Kong, in 2009. During 2008-2011, he was with the Department of Electrical Engineering, Columbia University, New York. He is currently an Associate Professor of Computer Science with Fudan University, Shanghai, China. His research interests include multimedia retrieval and computer vision. Dr. Jiang is an active participant of the Annual U.S. NIST TRECVID Evaluation and has designed a few top-performing video analytic systems over the years. His work has led to a best demo award from ACM Hong Kong, the second prize of ACM Multimedia Grand Challenge 2011, and a recognition by IBM T. J. Watson Research Center as one of ten "emerging leaders in multimedia" in 2009. He has served on the program committees of many international conferences and workshops and is a Guest Editor of a special issue on Multimedia Event Detection, Machine Vision and Applications, and a special issue on Socio-Mobile Media Analysis and Retrieval, IEEE Transactions on Multimedia.

Large-scale Multimedia Content Copy Detection in Mainstream Applications

Eric Zavesky
AT&T Labs Research, USA
ezavesky@research.att.com

1. Introduction

Multimedia content now occupies every facet of one's daily life: morning news alerts, multimedia email and texts, shared social content, and streamed mobile entertainment. Each area brings large volumes of multimedia content to everyone, but often the tools to manage this content are segregated and incomplete. Content-based Copy Detection is one technology near a commodity tipping point that will quickly spread through mainstream applications, just as the availability of the Viola-Jones AdaBoost face detection method has lead to intelligent cameras, social applications, and near convergence of face recognition. While there are calls for standards for content-based copy detection [1], and no hardware solutions currently exist, numerous consumer and enterprise applications are increasingly available, often as network-based services. This letter briefly discusses three blossoming topics for user-facing applications, active research areas, and questions that remain unresolved across the topic. While a number of services are described, each is bound by the goal of making the user experience more enjoyable and the technology more passive often by taking advantage of distributed network services.

2. Consumer Media Management

The proliferation of digital cameras and the desire to share personal content has given popularity to communication through sharing and social networks. Unfortunately, one critical capability currently underserved is the ability to easily organize and manage this consumer media. The most common methods for browsing users' photos, in decreasing availability, are: date, location, and textual tags. While interfaces can be constructed to leverage this weak metadata like timelines, flip-boards, trees, and grids, it is often the user that is responsible for creating a meaningful media organization. Alternatively, presenting initial organizations that are automatically computed (grouping photos of people or locations) or displaying extra metadata (proposed names or events of activities) to the user empowers media exploration in previously unavailable ways. A recent example of a dramatic shift in the user experience is the automated linkage of address book contacts to social services, which was adopted and is now expected on new devices.

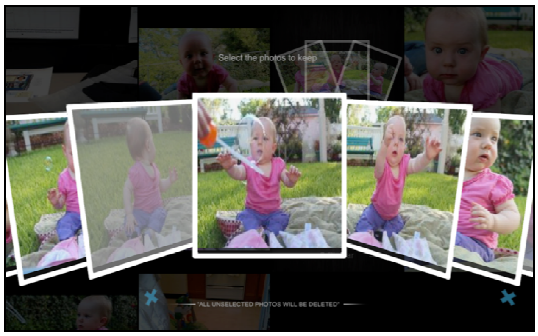


Figure 1: Similar photos from a single event automatically stacked by copy detection score

Leveraging automated processing in a similar way, methods developed at AT&T Labs in a consumer prototype system called VidCat provide "stacked" photos that are highly similar across albums from different times, poses, and capture angles [2]. Figure 1 illustrates a photo stack generated from copy detection scores allowing the user to quickly approve or subjectively prune a set for his or her best photo. Although temporal and geospatial constraints can also be applied, they are often unnecessary because of the initial high-precision copy detection. Like other vision-based detection and recognition systems, photo stacking of consumer content can be run as a distributed network service with no user input to eliminate frustrating browsing sessions of tens or hundreds of near-duplicate photos.

3. Blind Source Reconstruction

The high-quality visual precision that large-scale copy detection systems can achieve has also created a set of applications that can augment or replace missing data sources. Location sensors, while standard on mobile devices, still lack the ability to determine precise location in dense urban areas or indoor environments. These challenges can be overcome with copy detection systems evaluated over a very large-scale visual dataset to interactively guide the user and determine location and viewpoint with a minimal number of actions [3]. However as device manufacturers and applications are forced to operate in modes that conserve energy or reduce network traffic, some sensor data may become sparse or unreliable. Fortunately, visual copy detection allows cameras, sensors that demands continual operation to provide a high quality experience, to be

powered down if a large-scale dataset is available for a particular event. Bootstrapped by audio copy detection, one system synchronizes low quality consumer recordings from a public source to reconstruct full concert performances from many views [4]. An impressive capability of both of these systems is that with only content as user input, a highly accurate reconstruction of an event or location can be created with copy detection.

In a similarly blind source setting, a system was created in AT&T Labs that can automatically detect and enumerate repeated commercial segments over an arbitrary amount of time. With a coarse-to-fine search strategy, small pairs of visually matching video segments are promoted into sets of potential commercial segments. As more segments are pooled, precise temporal extents and frame-aligned offsets are computed. From half a year of HD content on one channel, 1745 manually identified commercials could be found with an F1 score of 92% with only 10 seconds of content. Such a system could be used in an incremental fashion to pinpoint these newly discovered video segments or to detect statistical trends and anomalies over an arbitrary dataset for surveillance, quality monitoring, and even automated asset management.

4. Product & Companion Linking

One may argue that application of product linking was viable with the first binary feature descriptor, but it is commonly acknowledged that publication of Lowe's visual SIFT descriptor [5] and Wang's audio "landmark" descriptor [6] actually bootstrapped services for product discovery from low quality queries. Now, these methods (or ones similar to them) are the heart of visual discovery services like SnapTell (now A9), Google Goggles, or Kooba and audio discovery services like Shazam or Audentify (now Autonomy, HP). These services allow users of mobile devices to quickly capture a photo or small audio segment and link to the object, book, movie, or song that they were experiencing.

Although these services may satisfy one aspect of a user's in-the-moment desires, new work has focused on live, synchronous experiences for users through companion devices. Here, services by IntoNow, Audible Magic, and Zeitera capture live streams, extract low-level visual and audio features and have them available in low-latency indexes for mobile search. Search results are repurposed for an improved user experience that allows access to new content for television viewers or a connection to a social network for immediate sharing of comments and opinions.

Both of these use cases require a large-scale solution that is capable of capture, intelligent segmentation, indexing, and retrieval. Various permutations of these networked technologies have been developed at AT&T Labs in prototype systems like CAE [7] and CAM [8]. The Content Analysis Engine (CAE) is a system that provides state-of-the-art shot boundary [9], speech transcription and alignment, and recognition tools to process broadcast and enterprise content [10]. Continuously online for over a decade, it has been adapted to index both professional and web-grade content with metadata including copy detection scores made available to user applications. The Content Augmenting Media (CAM) system harnesses output of the CAE and streams metadata to companion devices. This reverse-server model allows the copy detection (here companion linking) process can be run locally, thereby reducing up-stream network traffic (the most common bottleneck) and even allowing copy-detection "profiles" not available with traditional capture-and-query services.

5. Active Research & Evaluation

Although this letter has focused on mainstream applications, the topic of copy detection is active in many areas. Perhaps the seminal source for large-scale system evaluations is TRECVID, a set of challenging video-based tasks evaluated annually [11]. From 2008-2011, the content-based copy detection (CCD) task was evaluated that required systems to detect and identify original content segments after a battery of visual and audio transforms. Other evaluations have included object-based instance search (INS) and surveillance event detection (SED) that focus on object-level pattern queries.

A second branch of work focuses on alternate feature descriptors and indexing techniques. In most copy detection systems, the robustness, computational efficiency, and feature dimension count are each important aspects to consider for descriptor choice. Benchmarks that provide SIFT, SURF, and GIST, etc. descriptors are often published and commonly used to evaluate quantization and hashing techniques [12]. Other publications focused on specific applications like blind source reconstruction or product linking benchmarking collections of online or recaptured product images [13][14] and buildings [15].

6. Unique Problems

While several services have been created and deployed for various applications, a number of choices remain open for the system designer to choose. *Local or Networked?* Recent works have focused on optimization of local descriptors such chat they can be computed on mobile devices and transmitted to the query server with minimal network bandwidth [16]. Alternatively, systems evaluated in this letter, and

others commercially deployed, execute all feature extraction on the query server. Here, the tradeoff is network communication speed versus device portability and processing power; both of these variables continue to change, so no universal answer is available. *Context or Object?* For copy detection applications like product linking, one can choose to apply a semi- or fully-automated segmentation, like a salience-based method for object carving [17]. Systems evaluated in this letter and those that utilize global features utilize the entire scene context for a query, which may provide extra visual context that implicitly decreases false duplicate alarms. *Does Scale Generalize?* As indexed datasets grow in size, one may ask whether at very large scale can every possible query be generalized? A similar question was posed for *similar* images [18], but it is worth consideration when trying to solve event and gesture recognition problems like those described in section 5.

References

[1] "Compact Descriptors for Visual Search", MPEG Requirements Group, July 2010- Mar. 2011; http://mpeg.chiariglione.org/ working_documents.php.

[2] L. Begeja, E. Zavesky, Z. Liu, D. Gibbon, R. Gopalan, B. Shahraray "VidCat: An Image and Video Analysis Service for Personal Media Management". In SPIE, February 2013. *Submitted*

[3] F. Yu, R. Ji, S.-F. Chang. "Active Query Sensing for mobile location search." In ACM MM, November 2011.

[4] L. Kennedy, M. Naaman. "Less Talk, More Rock: Automated Organization of Community-Contributed Collections of Concert Videos." In WWW April 2009.

[5] D. Lowe, "Distinctive image features from scale-invariant keypoints," IJCV, 2004.

[6] A. Wang, "An Industrial-Strength Audio Search Algorithm." In ISMIR, October 2003.

[7] D. Gibbon, A. Basso, L. Begeja, Z. Liu, B. Renger, B. Shahraray, E. Zavesky. "Large-Scale Analysis for Interactive Media Consumption." In "TV Content Analysis, TV Content Analysis", CRC Press, 2012.

[8] "Content Augmenting Metadata." http://www.research.att.com/projects/Video/CAM/, August 2011.

[9] Z. Liu, D. Gibbon, E. Zavesky, B. Shahraray, P. Haffner. "A Fast, Comprehensive Shot Boundary Determination System". In ICME, July 2007.

[10] Z. Liu, T. Liu, D. Gibbon, B. Shahraray, "Effective and Scalable Video Copy Detection." In MIR, March 2010.

[11] A. Smeaton, P. Over, W. Kraaij, "Evaluation campaigns and TRECVid". In MIR October 2006.

[12] H. Je´gou, M. Douze, C. Schmid. "Product quantization for nearest neighbor search." In IEEE TPAMI, 2011.

[13] H. Kang, M. Hebert, A. Efros, T. Kanade. "Connecting Missing Links: Object Discovery from Sparse Observations using 5 Million Product Images", In ECCV, October 2012.

[14] V. Chandrasekhar, D. Chen, S. Tsai, N.-M. Cheung, H. Chen, G. Takacs, Y. Reznik, R. Vedantham, R. Grzeszczuk, J. Bach, and B. Girod, "The stanford mobile visual search data set". In MMSys 2011.

[15] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman. "Object retrieval with large vocabularies and fast spatial matching," In CVPR June 2007.

[16] B. Girod , V. Chandrasekhar , D. Chen , N. Cheung , R. Grzeszczuk , Y. Reznik, "Mobile Visual Search." In SPM, July 2011.

[17] M. Cheng, G..Zhang, N. Mitra, X. Huang, S.Hu. "Global contrast based salient region detection." In CVPR, June 2011.

[18] A. Torralba, R. Fergus, Y. Weiss. "Small codes and large image databases for recognition." In CVPR, June 2008.

Eric Zavesky is a Principal Member of Technical Staff at AT&T Labs Research. He received his B.S. degree in Electrical Engineering from The University of Texas in 2000 and his M.S. and Ph.D. degrees in Electrical Engineering from Columbia University in 2005 and 2010 respectively. After joining AT&T in 2009, he has collaborated on several projects to bring alternative query and retrieval representations to multimedia indexing systems including object-based query, biometric representations for personal authentication, and work to incorporate spatiotemporal information into near-duplicate copy detection. Eric holds 12 published or pending U.S. patents and 17 conference-level publications. His work in prior diverse fields includes semantic visual representations of content with low-latency, high-accuracy interactive search engines and start-ups focused on DVR with live-DVD authoring and low-latency, high quality audio synthesizers used in theatrical scores.

MMTC Officers

CHAIR

Jianwei Huang
The Chinese University of Hong Kong
China

STEERING COMMITTEE CHAIR

Pascal Frossard
EPFL, Switzerland

VICE CHAIRS

Kai Yang
Bell Labs, Alcatel-Lucent
USA

Chonggang Wang
InterDigital Communications
USA

Yonggang Wen
Nanyang Technological University
Singapore

Luigi Atzori
University of Cagliari
Italy

SECRETARY

Liang Zhou
Nanjing University of Posts and Telecommunications
China