RGB-D VIDEO CONTENT IDENTIFICATION

Honghai Yu [†*], *Pierre Moulin*[†*] *and Sujoy Roy*[‡]

[†]ECE Dept., University of Illinois at Urbana-Champaign, USA
[*]Advanced Digital Sciences Center, Singapore
[‡]Institute for Infocomm Research, A*STAR, Singapore

ABSTRACT

This paper proposes the first content identification (ID) system for depth video as well as a first hybrid content ID system for synchronized RGB and depth (RGB-D) video. The proposed systems are tested on a public RGB-D dataset. The hybrid system demonstrates significant performance gains over RGB-alone or depth-alone systems, while depth and RGB perform comparably. Moreover, a statistical interpretation of the hybrid system's superior performance is provided.

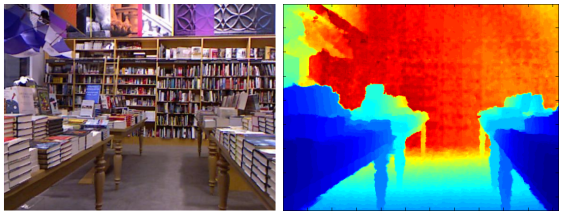
Index Terms— Content identification, fingerprinting, Kinect camera, depth video

1. INTRODUCTION

As people continuously search for better technologies to protect, manage and retrieval video content, content identification (ID) has received considerable attention from both academia and industry. Different from watermarking, which inserts an identifier into the the video content and thus changes the content, content ID extracts a signature (fingerprint) from the video content without changing it. A video fingerprint is a short summary of the video content that is robust to content-preserving distortions. The goal is then to match any query video to a database video by measuring the distance between the query fingerprint and the fingerprints in the database. Content ID can be used for filtering on file-sharing websites such as YouTube, advertisement tracking, broadcast monitoring, and law enforcement [1].

Many RGB video fingerprinting algorithms have been proposed and have demonstrated good performance [1]. Despite promising results, video content ID systems still face the limitation that video frames are 2-D projections of the 3-D world and depth information is lost. Fortunately, advances in sensing technology have now made it possible to equip videos with depth information. In particular, Xbox Kinect cameras are inexpensive ($\sim$ \$150) and output both RGB and depth videos as illustrated in Fig, 1 [2]. Kinect was originally designed for gaming, but soon found applications to various research problems in Signal Processing, Computer Vision, Robotic Navigation, and Computer Graphics. The application of Kinect to real-time human pose recognition won the best paper at the top computer vision conference CVPR [3] in 2011. We expect that RGB+depth (RGB-D) videos will become widespread in the future, and that databases such as [4, 5, 6, 7] will be commonplace. Since the combination of RGB and depth information is intrinsically more suitable than RGB alone or depth alone for representing scene content, the central goal of this paper is to investigate how depth information can help identify query videos. To the best of our knowledge, no depth content ID system currently exists.

This work was supported by NSF grant CCF 12-19145 and by AUIP fellowship from the A*STAR Graduate Academy, Singapore.



(a) (b)

Fig. 1. A RGB-D video frame from the NYU depth V2 dataset: (a) RGB image; (b) depth image. The combination of both modalities is referred to as RGB-D.

Notation: we follow the convention that uppercase letters represent random variables while lowercase letters represent particular realizations of these random variables. A vector is denoted by an underscore (e.g., $\underline{f}$) and a temporal sequence by a boldface letter (e.g., $\mathbf{f}$).

2. OVERVIEW OF A CONTENT ID SYSTEM

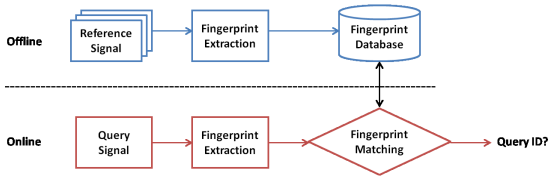


Fig. 2. Overview of a video content ID system

Fig. 2 shows a general content ID system applicable to most signal modalities, including RGB and depth video. The fingerprint database is built offline by extracting fingerprints from all reference signals. When a query signal (video) comes

in, its fingerprint is extracted and used as a query to the fingerprint database.

A key component of any content ID system is the fingerprint extraction algorithm which relies on signal processing primitives. In particular, learning algorithms that employ a variation of Adaboost to select filters and quantizers, such as Symmetric Pairwise Boosting (SPB) [8, 9] and regularized Adaboost [10], have demonstrated excellent content ID performance. The general framework is shown in Fig. 3 and is adopted in this paper.

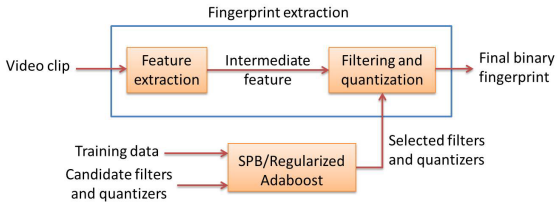


Fig. 3. Fingerprint Extraction Algorithm

The decoding metric in most content ID systems measures distance between fingerprints [8, 9, 11, 12]. If the distance is less than a predefined decision threshold τ, the fingerprint is declared as a match for the query. This is a *variable-size list decoder*: the number of matches could be 0, 1, 2 or more. Alternatively a single-output decoder might be used, returning only the index of the closest match. In this paper, Hamming distance metric with a list decoder is used in the experiments.

3. A DEPTH CONTENT ID SYSTEM

3.1. Depth Features

It can be difficult to extract good (i.e., robust and discriminative) fingerprints directly from raw video clips because of the difficulties of working with high dimensional data. In most video content ID systems, dimensionality reduction is applied before fingerprint extraction. We call the output of this step *intermediate features*.

Ideally, intermediate features should be sufficient statistics for the identification problem. It is unlikely any nontrivial such feature exists, and moreover the probability distribution of video data is not accurately known. Therefore, many heuristic RGB video features have been proposed and evaluated on some large datasets. This includes spatial features based on intensity change of a single image frame, temporal features based on a sequence of consecutive frames, color features computed in some color space, transformed domain features such as wavelet transform coefficients, or a combination of different types of features. The paper [1] provides an excellent review of RGB video features for content ID.

As Fig. 1 illustrates, depth images contain more homogeneous patches and fewer localized features, such as lines, edges and corners, than RGB images. If the intermediate feature $x \in \mathcal{X}$ consists of averages of homogeneous spatiotemporal patches of a depth video segment, x is approximately a sufficient statistic for the depth video segment. In practice, determining the number of homogeneous patches and their locations can be difficult and time consuming, and thus we propose the block mean depth (BMD) as intermediate features for depth video. First, each depth frame is divided into $N_r \times N_c$ blocks (N_r rows and N_c columns). The intermediate feature x at block $B_{r,c,t}$ in the r-th row, c-th column and t-th ($1 \le t \le T$) frame of a depth video segment is calculated as

$$x(r,c,t) = \frac{1}{|B_{r,c,t}|} \sum_{(i,j) \in B_{r,c,t}} d(i,j,t), \qquad (1)$$

where $|\cdot|$ denotes set cardinality, and $d(i,j,t)$ is the depth value at coordinates (i,j) in the t-th frame. Hence, the feature space $\mathcal{X} = \mathbb{R}^{N_r \times N_c \times T}$. The averaging operation in the BMD feature makes it ralatively robust to the acquisition noise in depth frames. Another motivation is that block mean luminance (BML), the counterpart of BMD for RGB video, has demonstrated excellent robustness to lossy compression, frame resizing and frame rate change in [9]. We have observed the same nearly perfect performance of BMD for depth video, and thus we have considered more challenging degradations such as cropping and rotation in our experiments to test BMD's robustness and discriminative power.

3.2. Filter and Quantizer Selection

Based on the extracted intermediate features, the filter and quantizer selection algorithms *symmetric pairwise boosting* (SPB) [9] and *regularized Adaboost* [10] operate as follows. A training set $\mathcal{T} \triangleq \{(x_t, y_t, z_t) \in \mathcal{X}^2 \times \{\pm 1\}, t \in \mathcal{T}\}$ is comprised of a subset $\mathcal{T}_+$ of $|\mathcal{T}|/2$ *matching pairs* and a subset $\mathcal{T}_-$ of $|\mathcal{T}|/2$ *nonmatching pairs*, where a pair $(x_t, y_t) \in \mathcal{X}^2$ is said to be matching if the second signal is a distorted version of the first, and nonmatching if the two signals are independent. The binary variable (label) z_t is equal to 1 (resp. -1) if (x_t, y_t) is matching (resp. nonmatching). Define a set of J weak classifiers $h_j : \mathcal{X}^2 \to \{\pm 1\}, 1 \le j \le J$, as

$$h_j(x,y) = \begin{cases} +1 & \text{if } \phi_j(x) = \phi_j(y) \\ -1 & \text{otherwise} \end{cases} \qquad (2)$$

where $\phi_j(x) = Q_j(\lambda_j(x))$ is parameterized by a filter $\lambda_j : \mathcal{X} \to \mathbb{R}$ and a quantizer $Q_j : \mathbb{R} \to \mathcal{A}$. Denote by $\mathcal{H}$ the class of feasible classifiers (indexed by the choice of filters and quantizers).

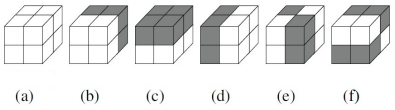


Fig. 4. 3-D Haar-like filters [9]: (a) spatio-temporal average, (b) temporal difference, (c,d) spatial difference, and (e,f) spatio-temporal difference.

A popular family of filters is the Haar-like filters used in [8, 9, 13] which are easy to compute and rich enough to describe perceptually significant visual features. The filter outputs for the 3-D Haar-like filters in [9] are the average difference between values in light and dark regions shown in Fig. 4.

To reduce the computational complexity of the training, a limited number of candidate quantizers are evaluated.

The SPB algorithm is an adaptation of the well-known Adaboost classification algorithm [9], while the regularized Adaboost algorithm used a regularizer to effectively eliminate those classifiers that generate highly correlated fingerprints from the candidate pool $\mathcal{H}$ [10]. Upon completion of the algorithm, both algorithms would output the *boosted classifier* $h_{\mathrm{B}}(x,y) \triangleq \mathrm{sgn}\left[\sum_{1\le j\le J}\alpha_j h_j(x,y)\right]$, from which only the filter λ_j and quantizer Q_j associated with each h_j are used to produce the fingerprints.

Given a video signal $\mathbf{x} = x_1,\ldots,x_L \in \mathcal{X}^L$ consisting of L video segments, the fingerprint is obtained as an array $\mathbf{f} = \{f_{ij}, 1\le i\le L, 1\le j\le J\}$ where i denotes time and j classifier index, and $f_{ij} = \phi_j(x_i)$. We also use the notation $\underline{f} = \{f_j, 1\le j\le J\}$ for the subfingerprint associated with a given video segment.

4. RGB-D CONTENT ID SYSTEM

As illustrated in Fig. 1, most humans can reasonably infer the depth information from the corresponding RGB image. A mathematical algorithm has recently been developed to estimate depth information from a single RGB image [14]. However, this kind of inference requires global image processing, and local features such as block mean luminance (for RBG) and block mean depth are more likely to be independent. Thus we can build a hybrid system to harvest the diversity gain, when both RGB and depth are available in video signals.

Fig. 5 illustrates our RGB-D fingerprint code design. We train half of the filters and quantizers from RGB intermediate features, and the other half from depth intermediate features. Thus for each RGB-D video, half of the fingerprint is generated from RGB and half from depth. Then filtering and quantization are applied, and the combined final fingerprint is used to identify the video in the hybrid system.

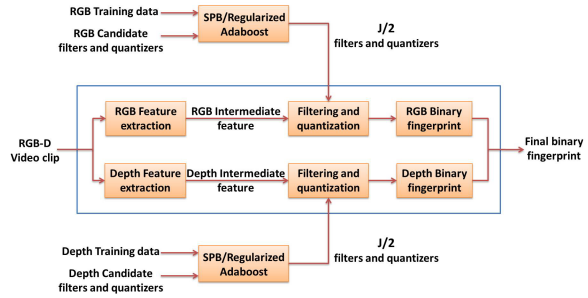


Fig. 5. Fingerprint extraction for a hybrid system.

5. PERFORMANCE EVALUATION

5.1. Experimental Setup

The NYU Depth V2 dataset captures comprehensive indoor environments and used in our experiments. It is comprised of 464 indoor scenes taken from three cities. Each scene is recorded as a short RGB-D video. We use 115 videos for training and another 115 videos for testing. The training set and testing set are randomly selected and contain a roughly equal number of scenes from each scene type. The training data includes 16,000 matching and 16,000 nonmatching pairs ($|\mathcal{T}| = 32,000$) of sequences of intermediate features from 10 consecutive synchronized RGB and depth frames. The matching pairs are generated from the video distortions illustrated in Fig. 6: 50% cropping, vertical mirroring, frame rotation of 15 degrees and frame shifting downward and left by 100 pixels. We consider geometric distortions only as they represent the most challenging video distortions to detect. Both SPB and regularized Adaboost work nearly perfectly for simple distortions such as lossy compression, resizing, and frame rate change for both RGB [9, 10] and depth video. The nonmatching pairs are generated from intermediate feature sequences extracted from different RGB-D videos.

We adopt the same video normalization as in [9]. Before extraction of intermediate features, both RGB and depth videos are resampled at 10 frames per second, converted to grayscale (RGB only) and resized to QVGA (320x240). These preprocessing steps aim to make the fingerprinting algorithm robust to frame rate change, color variation, and frame resizing. After preprocessing, block mean luminance (BML) and block mean depth (BMD) are extracted from RGB and depth video clips on 36 ($N_r = 4, N_c = 9$) blocks per frame. The temporal length of the intermediate features is 1 second, and the query length is 5 seconds with overlapping factor of $9/10$. We train $J = 16$ classifiers each for RGB and depth. The first 8 classifiers from RGB and depth are combined to generate hybrid fingerprints. Each filter output is quantized into 4 levels. Hence our query fingerprint is 1312 bits long. For regularized Adaboost, we use a regularization parameter [10] of $\gamma = 0.2$ for RGB and $\gamma = 0.1$ for depth.

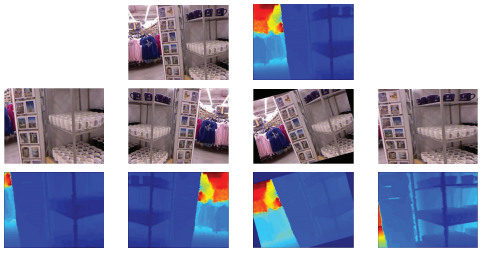


Fig. 6. Sample distorted images. Top row: Original RGB and depth images. Bottom two rows: Distorted RGB and depth images. Distortions from left to right are: cropping of 50%, vertical mirroring, rotation of 15 degree and shift downward and left by 100 pixels.

5.2. Experimental Results

As shown in Fig. 7, the hybrid system outperforms the RGB-alone and depth-alone systems for all the considered distortions, irrespective of the fingerprinting algorithms used, while RGB and depth perform comparably for cropping and shitting. Moreover, regularized Adaboost performs significantly better than SPB, irrespective of the modalities used. For the image rotation distortion of Fig. 7g, hybrid system's perfor-

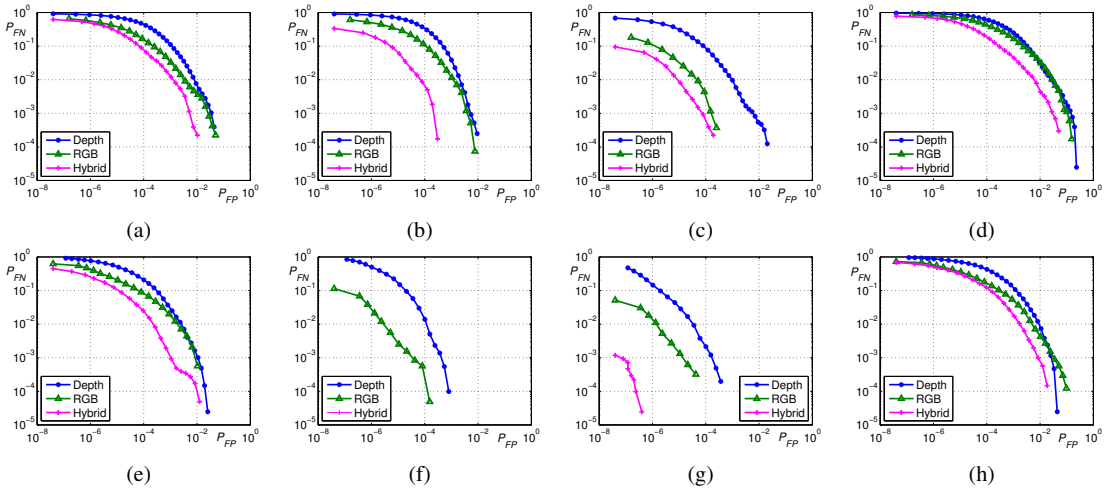


Fig. 7. First row: ROC curves for SPB under distortions: (a) cropping; (b) vertical mirroring; (c) rotation; (d) shift.
Second row: ROC curves for regularized Adaboost under distortions: (e) cropping; (f) vertical mirroring; (g) rotation; (h) shift.

mance gain is in orders of magnitude. Even more stunning is the result for the distortion of vertical mirroring in Fig. 7f, where false negative rate is zero for all possible values of false positive rates based on a simulation of 25,073,193 nonmatching query pairs.

5.3. Statistical Interpretation

In general, the superiority of regularized Adaboost over SPB stems from its ability to select more independent features [10]. A similar phenomenon applies here, especially when we consider RGB features and depth features.

We first define the within-modality correlation of two filters λ_j and λ_k $(1 \leq j, k \leq J)$ as

$$R^m(j,k) = \frac{\mathbb{E}\left[(\lambda_j^m(X^m) - \mu_j^m)(\lambda_k^m(X^m) - \mu_k^m)\right]}{\sigma_j^m \sigma_k^m}, \quad (3)$$

where $m \in \{\text{RGB}, \text{D}\}$ denotes RGB and depth (D) respectively, X^m is the intermediate feature of one segment from the corresponding modality, μ_j^m and σ_j^m are the mean and standard deviation of $\lambda_j^m(X^m)$. We also define the between-modality correlation

$$R^{\text{RGB-D}}(j,k) = \frac{\mathbb{E}\left[(\lambda_j^{\text{RGB}}(X^{\text{RGB}}) - \mu_j^{\text{RGB}})(\lambda_k^{\text{D}}(X^{\text{D}}) - \mu_k^{\text{D}})\right]}{\sigma_j^{\text{RGB}} \sigma_k^{\text{D}}}, \quad (4)$$

the average absolute within-modality correlation,

$$\overline{R}^m = \frac{2}{J^2 - J} \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} |R^m(j,k)|, \quad (5)$$

and the average absolute between-modality correlation,

$$\overline{R}^{\text{RGB-D}} = \frac{1}{J^2} \sum_{j=1}^{J} \sum_{k=1}^{J} |R^{\text{RGB-D}}(j,k)|, \quad (6)$$

of these filters. The expectations are estimated from the training dataset, and their values are shown in Table 1. The average between-modality correlation is almost an order of magnitude smaller than the average within-modality correlation.

	$\overline{R}^{\text{RGB}}$	$\overline{R}^{\text{D}}$	$\overline{R}^{\text{RGB-D}}$
SPB	0.1496	0.1025	0.0208
Regularized Adaboost	0.2303	0.1705	0.0199

Table 1. Average within-modality and between-modality correlations.

We also show in Fig. 8 the distributions of Hamming distance for matching and nonmatching pairs under the vertical mirroring distortion (other distortions exhibit the same trend). The better histogram separation of the second row is consistent with regularized Adaboost's superior content ID performance in Fig. 7f over SPB in Fig. 7b. The clear improvement in histogram separation from the left two columns (single modality) to the last column (RGB-D) is consistent with the better ROC curves of hybrid systems in Fig. 7b and Fig. 7f. Overall, a hybrid system based on regularized Adaboost performs significantly better than the other systems we considered.

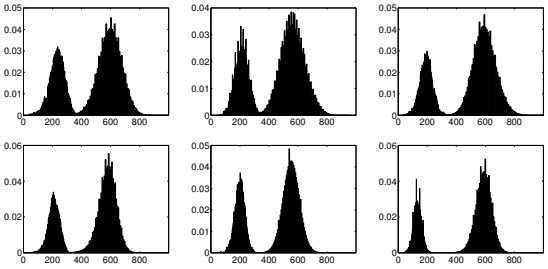


Fig. 8. Distributions of Hamming distance for matching and nonmatching pairs for the vertical mirroring distortion. First row from left to right are SPB for: depth, RGB, and RGB-D. Second row from left to right are regularized Adaboost for: depth, RGB, and RGB-D.

6. REFERENCES

[1] J. Lu, "Video fingerprinting for copy identification: from research to industry applications," in *SPIE Electric Imaging Symposium on Media Forensics and Security I*, San Jose, CA, USA, 2009.

[2] "http://www.xbox.com/en-US/KINECT," .

[3] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, June 2011.

[4] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proceedings of the European Conference on Computer Vision*, Firenze, Italy, October 2012.

[5] C. Wolf, J. Mille, L.E., Lombardi, O. Celiktutan, M. Jiu, M. Baccouche, E. Dellandrea, C. E. Bichot, C. Garcia, and B. Sankur, "The LIRIS human activities dataset and the ICPR 2012 human activities recognition and localization competition," Tech. Rep., LIRIS Laboratory, 2012.

[6] K. Lai, L. Bo, X. Ren, and D. Fox, "A Large-Scale Hierarchical Multi-View RGB-D Object Dataset," in *IEEE International Conference on on Robotics and Automation*, Shanghai, China, May 2011.

[7] B. Ni, G. Wang, and P. Moulin, "RGBD-HuDaAct: A color-depth video database for human daily activity recognition," in *IEEE Workshop on Consumer Depth Cameras for Computer Vision in conjunction with ICCV*, 2011.

[8] D. Jang, C. D. Yoo, S. Lee, S. Kim, and T. Kalker, "Pairwise boosted audio fingerprint," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 4, pp. 995–1004, Dec. 2009.

[9] S. Lee, C. D. Yoo, and T. Kalker, "Robust video fingerprinting based on symmetric pairwise boosting," *IEEE Trans. Circ. and Sys. for Video Technol.*, vol. 19, no. 9, pp. 1379–1388, Sept. 2009.

[10] H. Yu and P. Moulin, "Regularized Adaboost for content identification," *Accepted to ICASSP*, 2013, Available from `http://www.ifp.illinois.edu/~yu75/RegularizedAdaboost.pdf`.

[11] P. Moulin, "Statistical modeling and analysis of content identification," in *Information Theory and Applications Workshop*, San Diego, CA, February 2010.

[12] R. Naini and P. Moulin, "Model-based decoding metrics for content identification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, March 2012, pp. 1829 –1832.

[13] Y. Ke, D. Hoiem, and R. Sukthankar, "Computer vision for music identification," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Washington, DC, USA, 2005, pp. 597–604.

[14] A. Saxena, S. H. Chung, and A. Y. Ng, "3-D depth reconstruction from a single still image," *International Journal of Computer Vision (IJCV)*, vol. 76, 2007.