Deep Video Hashing

Venice Erin Liong, Jiwen Lu, *Senior Member, IEEE*, Yap-Peng Tan, *Senior Member, IEEE*, and Jie Zhou, *Senior Member, IEEE*

***Abstract*—In this work, we propose a deep video hashing (DVH) method for scalable video search. Unlike most existing video hashing methods that first extract features for each single frame and then use conventional image hashing techniques, our DVH learns binary codes for the entire video with a deep learning framework so that both the temporal and discriminative information can be well exploited. Specifically, we fuse the temporal information across different frames within each video to learn the feature representation under two criteria: the distance between a feature pair obtained at the top layer is small if they are from the same class, and large if they are from different classes; and the quantization loss between the real-valued features and the binary codes is minimized. We exploit different deep architectures to utilize spatial-temporal information in different manners and compare them with single-frame-based deep models and state-of-the-art image hashing methods. Experimental results demonstrate the effectiveness of our proposed method.**

***Index Terms*—Deep learning, scalable video search, video hashing.**

I. Introduction

SCALABLE visual search has gained large interests in computer vision, especially with the increasing amount of visual data which are available over the internet in recent years. This task aims to retrieve the most relevant visual content from a database for a query sample, in an accurate and efficient manner. In contrast to images, videos provide diverse and complex visual patterns consisting of low-level visual content in each frame as well as high-level structured content across frames [1]–[5], which makes video search more challenging than image search. Moreover, each video may have a number of image frames which leads to exhaustive computation between frames, which is impractical for large-scale video databases. Hence, how to develop a framework for scalable video search where discriminative information from videos can be well exploited remains an important problem in visual search.

A key topic in visual search is *learning-based hashing*, which transforms high-dimensional feature vectors to compact binary codes, by preserving visual content using statistical inference. However, most current works in scalable visual search focus on image-based retrieval [6]–[14] or text-image/image-text retrieval [15]–[19]. To our best knowledge, there are only few works that present an efficient framework for video hashing. Hence, it is desirable to make better use of hashing methods to exploit the spatial-temporal information of videos.

Current video hashing methods are mainly applied in two multimedia computing applications. The first is the near-duplicate search where conventional hashing techniques [20]–[22] were used to identify duplicate videos efficiently [23]–[25]. The second is content-based video retrieval [5], [11], [26], [27] which retrieves the most semantically similar videos from a database for a given query video. In this work, we focus on the latter one. Video hashing for content-based retrieval can be divided into three categories. The first extracts a single representative feature vector, and then performs hashing. The second treats each frame as an image and performs image hashing first such that the resulting frame-frame hamming distances of two videos are then combined through averaging. The last selects several representative frames and employs image hashing on these selected frames. While these frameworks are straightforward, they cannot exploit the temporal information of videos in the learning stage. Furthermore, hashing image frames individually is computationally expensive considering that each video would usually have a minimum of 30–50 frames. Hence, it is desirable to present a video hashing framework which learns strong features by utilizing the temporal information, and in turn, also minimizes the hamming distance computation.

To address these challenges, we propose a Deep Video Hashing (DVH) method for scalable video search where we explore different CNN-based architectures. Since Convolutional Neural Networks (CNNs) have shown promising performance in several computer vision tasks due to their strong feature representation capability, we also employ CNN in our video hashing framework. The basic idea of our approach is shown in Fig. 1. Specifically, we build an end-to-end deep CNN learning framework which utilizes spatio-temporal information after the

Manuscript received August 2, 2016; revised December 7, 2016; accepted December 19, 2016. Date of publication December 26, 2016; date of current version May 13, 2017. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001001, in part by the National Natural Science Foundation of China under Grant 61672306, Grant 61225008, Grant 61572271, Grant 61527808, Grant 61373074, and Grant 61373090, in part by the National 1000 Young Talents Plan Program, in part by the MSRA Collaborative Research Program, in part by the National Basic Research Program of China under Grant 2014CB349304, in part by the Ministry of Education of China under Grant 20120002110033, and in part by the Tsinghua University Initiative Scientific Research Program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jingdong Wang. *(Corresponding author: Jiwen Lu.)*

V. E. Liong is with the Interdisciplinary Graduate School, Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore 639798 (e-mail: veniceer001@e.ntu.edu.sg).

J. Lu and J. Zhou are with the Department of Automation, State Key Laboratory of Intelligent Technologies and Systems, Tsinghua University, Beijing 100084, China, and also with the Tsinghua National Laboratory for Information Science and Technology, Beijing 100084, China (e-mail: lujiwen@tsinghua.edu.cn; jzhou@tsinghua.edu.cn).

Y.-P. Tan is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: eyptan@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2645404

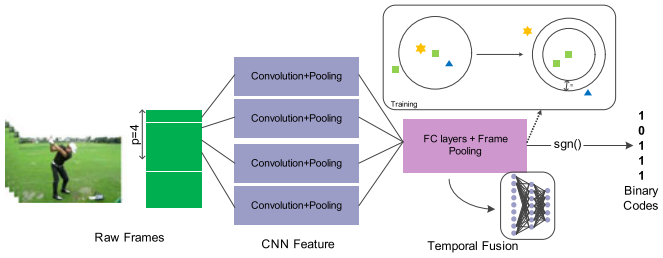

Fig. 1. Basic idea of our proposed DVH approach. We pass a set of successive image frames with a predefined frame number (in this figure, $p = 4$) to the convolutional and pooling layers to obtain frame-wise CNN feature representation. Then, we perform temporal fusion in the fully connected layers. At the final stage of our deep network, we apply the $\mathrm{sgn}(\cdot)$ in the activation outputs and obtain the compact binary codes. At the training stage, the network parameters are learned using back-propagation with a large-margin cost function such that video features that contain similar semantics/label information are close to each other, while video features that are dissimilar are far apart as possible, and a quantization loss criterion such that the real feature codes and binary codes are similar as much as possible.

stacked convolutional-pooling layers to extract representative video features, and obtain compact binary codes. We discriminatively train our model with a Siamese network, by maximizing the inter-class distance and minimizing the intra-class distance of the video feature pairs, as well as minimizing the quantization loss between real-value codes and binary codes. We exploit different spatial-temporal feature pooling architectures and compare them to single-frame CNN architectures as well as state-of-the-art image-based hashing methods. Experimental results on two video datasets show the effectiveness of our proposed approach.

Our contributions are summarized as follows:

1) We propose a deep learning-based hashing method called deep video hashing (DVH) which learns a deep network to exploit the discriminative and temporal information of videos in order to represent each video with meaningful binary codes.
2) We conduct extensive scalable video search experiments on two video datasets to demonstrate the efficacy of our proposed method. We exploit different frame fusion architectures and compare them to several baseline network architectures, state-of-the-art single frame hashing methods, and existing video hashing methods.

II. Related Work

In this section, we briefly review three related topics: 1) learning-based hashing, 2) video hashing, and 3) deep learning for video analysis.

A. Learning-Based Hashing

Several learning-based hashing methods such as subspace models [6], [28], manifold models [29], [30], and kernel models [31], [32] have been exploited in the literature. These methods can be classified into two classes: *unsupervised* [6], [7], [20], [33], [34] and *supervised* [9], [28], [30], [31], [35]–[37]. The first category does not require label information. For example, Gong *et al.* [6] proposed a PCA-ITQ method which learns hashing functions by first performing PCA to maximize the variance of the hash bits and then learning a rotation matrix to minimize the quantization loss. Liu *et al.* [38] proposed an Anchor Graph Hashing (AGH) method by using the concept of *anchors* to identify the similarity between features. For the second category, class-wise label information is utilized. For example, Gong and Lazebnik [6] extended the PCA-ITQ to CCA-ITQ which first performs CCA to maximize the correlation between semantically similar features, and then minimizes the quantization loss. Liu *et al.* [31] employed the Kernel Supervised Hashing (KSH) method to perform kernel mapping and utilize supervised information through minimizing the distance of similar pairs and maximizing the distance of dissimilar pairs. Lin *et al.* [37] proposed a FastHash method to learn the binary codes through Graph Cuts and a greedy boosted decision tree framework. While these methods are mostly nonlinear hashing techniques, only a few works have used deep learning techniques to perform end-to-end nonlinear mapping [39]–[43]. Furthermore, these methods are specifically developed for large-scale image retrieval. In contrast, our work focuses on using statistical knowledge for scalable video search.

B. Video Hashing

Previous works on video hashing were mostly used for near-duplicate video retrieval tasks and several of them focused on video feature representation rather than learning the hashing functions [5], [23]–[25], [44]. For example, Song *et al.* [23] introduced a multiple feature hashing method which utilizes multiple features and extracts different local structures to obtain efficient binary codes. Douze *et al.* [24] extracted representative spatial-temporal features from images and used conventional hashing methods to obtain binary codes. There are only a few video hashing methods proposed for content-based retrieval. For example, Cao *et al.* [27] proposed a submodular hashing framework which selects relevant frames from videos to learn the hashing functions for efficient video search. However, it did not really learn a video-based hashing function by using statistical knowledge. Ye *et al.* [26] proposed a supervised structural learning framework which exploits the temporal consistency to learn the linear hashing functions. However, it only learns a linear projection which may not truly capture the nonlinear nature of video data. Li *et al.* [11], [45] proposed a hashing model across the Euclidean space and the Riemannian manifold, which learns hashing functions based on the kernel max-margin framework for face video retrieval. However, their work represented videos with a single feature representation (covariance matrix), which may not fully exploit the spatio-temporal information in videos.

C. Deep Learning for Video Analysis

Deep learning techniques have shown great success in various computer vision tasks such as in image recognition [46], [47], scene labeling [48], [49], and pedestrian detection [50]. While a number of deep learning methods have also been proposed for video analysis, most of them focused on video action recognition [51]–[54], video classification [55]–[57] and event detection [58], [59]. For example, Ji *et al.* [51] introduced a 3D

Convolution Neural Networks approach which considers spatial and temporal information for action recognition. Karpathy *et al.* [55] presented an extensive evaluation of different deep architectures for large-scale video classification where they introduced different spatio-temporal convolutions based on how the frames are fused in the network. Ng *et al.* [56] exploited different feature fusion methods after the stacked convolution and pooling layers and investigated the Long Short Term Memory (LSTM) networks for video classification. Xu *et al.* [58] explored pooling and encoding methods to combine frame-level features for event detection. Simonyan and Zisserman [53] implemented a two stream convolution network for action recognition in videos to model the spatial and temporal data individually, and fuse the scores together. To our knowledge, nobody has investigated deep architectures for video hashing.

III. Proposed Approach

In this section, we first present the basic idea of the learning-based video hashing model, then describe our deep video hashing (DVH) method and its implementation details.

A. Learning-Based Video Hashing

Let $\mathcal{X} = \{\mathbf{X}_i, y_i\}_{i=1}^{M}$ be a collection of M videos where $\mathbf{X}_i = [\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \ldots, \mathbf{x}_{i,f_i}] \in \mathbb{R}^{d \times f_i}$ is the ith video with successive f_i frames, y_i is the label information, and $\mathbf{x}_{ij} \in \mathbb{R}^d$ is the jth image feature frame of the video $\mathbf{X}_i$ with a feature length of d. The objective of learning-based video hashing is to learn K hash functions to project each video into a single or multiple K-bit compact binary vectors as follows:

$$\mathcal{F}_{\mathbf{X}_i} : \mathbb{R}^{d \times f_i} \rightarrow \{-1, 1\}^{K \times g_i} \tag{1}$$

where $g_i \in [1, f_i]$.[1]

To obtain hashing functions, we learn a linear projection matrix, $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_K] \in \mathbb{R}^{d \times K}$ to map the video features into compact binary codes. To obtain a single binary vector for the ith video, $\mathbf{b}_i \in \{-1, 1\}^{K \times 1}$, we first extract a compact single feature representation for the ith video defined as $\tilde{\mathbf{x}}_i$, and then project it linearly as follows:

$$\mathbf{b}_i = \text{sgn}(\mathbf{W}^\top \tilde{\mathbf{x}}_i). \tag{2}$$

However, it is difficult to represent a whole video as a single binary code without losing significant amount of information. Hence, a single video can be represented into multiple binary vectors by treating each frame as an image feature and performing image-based hashing. To obtain multiple binary vectors for the ith video, $\mathbf{B}_i \in \{-1, 1\}^{K \times f_i}$, we compute the frame-wise feature representation as follows:

$$\mathbf{B}_i = \text{sgn}(\mathbf{W}^\top \mathbf{X}_i). \tag{3}$$

These conventional learning-based video hashing methods make use of hand-crafted single representation features [11], [45] and/or use a single linear projection [26], which may not effectively capture the nonlinear relationship of video representations and cannot exploit temporal information present in videos.

B. Deep Video Hashing

Our work employs a deep learning model to learn several nonlinear projections to obtain compact binary codes, where both the discriminative and temporal information of videos are exploited in an end-to-end learning framework. By doing so, we are able to learn powerful video representations in a spatial-temporal level. Unlike previous video hashing methods which either learn hashing functions from a single video feature representation or frame-by-frame, we process a set of successive frames to obtain a single binary vector which leads to a fewer number ($g_i < f_i$) of binary codes to represent the ith video. Therefore, we are able to minimize the number of binary codes to represent each video but still extract significant information as much as possible.

As shown in Fig. 1, given a fixed frame size p, we have a set of image frames $\mathbf{I}_u \in \mathbb{R}^{p \times h \times w \times 3}$ that is passed through a series of convolution and pooling layers with fully connected layers at the end. By letting $s(\cdot)$ be the output at the last fully connected layer where it contains K nodes, the binary code for the set of image frames of the ith video is computed as follows:

$$\mathbf{b}_u = \text{sgn}(s(\mathbf{I}_u)). \tag{4}$$

There are two important intuitions for our DVH model: 1) By performing several nonlinear transformations with a discriminative criterion, more robust visual representation can be obtained. While kernel-based models can provide explicit nonlinear mappings, pre-defined kernel functions cannot well capture the nonlinearity of samples; 2) By performing the temporal pooling, we can implicitly exploit the relevant frames and extract a balance of global and local information from video frames. By doing so, the noisy frames which may degrade the quality of the binary codes can be implicitly ignored. We discuss how to exploit both the temporal and discriminative information in our deep architecture as follows:

1) Temporal Information: Since videos typically represent motion-based features across time, it is necessary to consider temporal information to fully obtain video-level representation. In order to exploit temporal information in deep networks, we perform pooling operations across frames between the fully-connected layers. In our work, we perform fusion of temporal information through average pooling.[2] We describe three deep networks with various feature pooling architectures:

Early fusion: The early fusion architecture first passes image frames through the convolution and pooling layers and then fuses the information at the first fully connected layer immediately, as shown in Fig. 2(a).

Late fusion: The late fusion architecture first passes image frames through the convolution and pooling layers up to the other fully-connected layers and then fuses the information at the last fully connected layer, as shown in Fig. 2(b).

[1]During learning, the binary codes are set to $\{-1,1\}$ to ensure proper centering, but they can then be set to $\{0,1\}$ during retrieval.

[2]We found that max pooling does not give very representative information compared to average pooling.

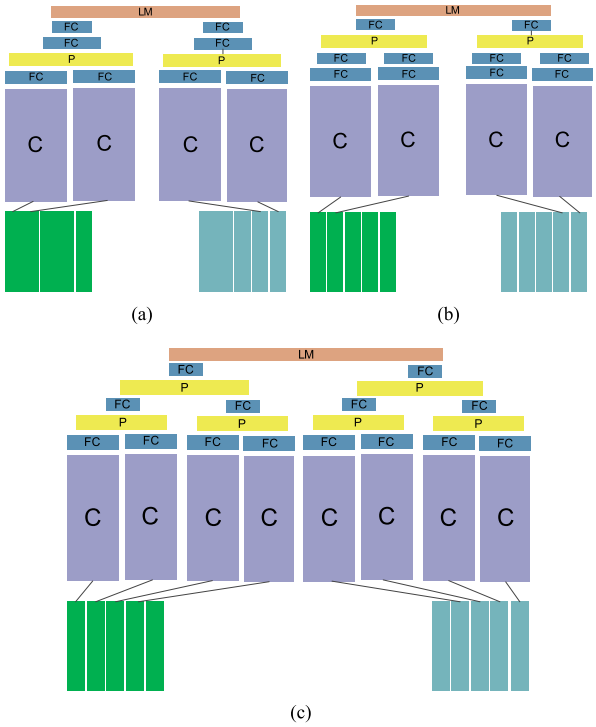

Fig. 2. Different DVH architectures with frame fusion, where the vertical bars represent the raw image frames and each color represents a single video, "C" represents the stacked convolution and pooling layers, "FC" represents the fully-connected layers, "P" represents the temporal pooling layers, and "LM" represents the large-margin cost function, respectively. (a) Early fusion. (b) Late fusion. (c) Slow fusion.

Slow fusion: The slow fusion architecture is a balance of the early and late fusion. Image frames are passed through the convolution and pooling layers and then fused in a hierarchical manner such that smaller temporal windows are used as it approaches the top layer. In this work, a two-stage fusion strategy is implemented. Fig. 2(c) details the architecture of this fusion strategy.

2) Discriminative and Binary Information: To learn the parameters in the deep network discriminatively, we employ the Siamese network [60] with a large-margin learning framework rather than the conventional contrastive divergence criterion [61]. Specifically, we present a new formulation which consists of two new objective criterions for binary code learning. The first objective performs discriminative learning. Specifically, given two sets of image frames, $\mathbf{I}_u$ and $\mathbf{I}_v$, we minimize the intra-class variation and maximize the inter-class variation of the binary feature representation at the top layer of these two networks, simultaneously. Given their Hamming distance $d_{u,v}(\mathbf{b}_u, \mathbf{b}_v)$ at the top layer, where $\mathbf{b}_u = \text{sign}(s(\mathbf{I}_u))$ and $\mathbf{b}_v = \text{sign}(s(\mathbf{I}_v))$, we expect that $d_{u,v}$ is small if u and v are the same class, and large if they are from different classes, which is formulated as the following constraints:

$$d_{u,v}(\mathbf{b}_u, \mathbf{b}_v) \leq \theta_1, \text{if } y_u = y_v \tag{5}$$

$$d_{u,v}(\mathbf{b}_u, \mathbf{b}_v) \geq \theta_2, \text{if } y_u \neq y_v \tag{6}$$

where θ_1 and θ_2 are the small and large thresholds, respectively.

By combining (5) and (6), we have the following formulas:

$$\delta_{u,v}\left(\theta - d_{u,v}(\mathbf{b}_u, \mathbf{b}_v)\right) > 1 \tag{7}$$

where $\theta_1 = \theta - 1$ and $\theta_2 = \theta + 1$, and $\delta_{u,v} = 1$ means that u and v are from the same class, and $\delta_{u,v} = -1$ indicates that they are from different classes.

This leads to the following objective function:

$$J_1 = f(1 - \delta_{u,v}(\theta - d_{u,v}(\mathbf{b}_u, \mathbf{b}_v))) \tag{8}$$

where $f(z)$ is a generalized logistic loss function which is a smooth approximation of the hinge loss function: $z = \max(z, 0)$, and defined as follows:

$$f(z) = \frac{1}{\rho}\log(1 + \exp(\rho z)) \tag{9}$$

where ρ is the sharpness parameter set to 10 and θ is the threshold parameter set to $K/4$.

The second objective is to ensure efficient binary codes by minimizing the quantization loss [6] between real-valued codes and binary codes as follows:

$$J_2 = \|s(\mathbf{I}_u) - \mathbf{b}_u\|_F^2 + \|s(\mathbf{I}_v) - \mathbf{b}_v\|_F^2. \tag{10}$$

Hence, the final objective function for our DVH is formulated as

$$\begin{aligned} \min_{\mathbf{b}_v, \mathbf{b}_v} J &= J_1 + \lambda J_2 \\ &= f(1 - \delta_{u,v}(\theta - d_{u,v}(\mathbf{b}_u, \mathbf{b}_v))) \\ &\quad + \lambda(\|s(\mathbf{I}_u) - \mathbf{b}_u\|_F^2 + \|s(\mathbf{I}_v) - \mathbf{b}_v\|_F^2) \end{aligned} \tag{11}$$

where J_1 exploits the discriminative information, J_2 minimizes the quantization loss, and λ is a constant parameter which balances the two criterions.

We use the standard mini-batch gradient descent and back-propagation to solve the optimization problem. We first relax the binary constraints in the first objective and use the signed magnitude real codes during back-propagation. Given $\mathbf{h}_u = s(\mathbf{I}_u)$ and $\mathbf{h}_v = s(\mathbf{I}_v)$ are the real-value code values from the top layer, the back-propagation is implemented first by taking the derivative of J with respect to $\mathbf{h}_u$ and $\mathbf{h}_v$

$$\frac{\partial J}{\mathbf{h}_u} = f'(z)\delta_{u,v}(\mathbf{h}_u - \mathbf{h}_v) + \lambda(\mathbf{h}_u - \mathbf{b}_u) \tag{12}$$

$$\frac{\partial J}{\mathbf{h}_v} = f'(z)\delta_{u,v}(\mathbf{h}_v - \mathbf{h}_u) + \lambda(\mathbf{h}_v - \mathbf{b}_v). \tag{13}$$

These gradients are then back-propagated to update the weights at the earlier layers.

C. Extracting Binary Codes From One Video

Since the number of frames are usually different for different videos, we extract multiple binary codes for each video by using our DVH approach through a set of consecutive image frames in the video with a specific stride. Fig. 3 shows the hashing procedure for one video by using our model. The over-all hamming distance, D_H, for a pair of videos are then obtained by getting the average of the hamming distances, d_H, for each pair

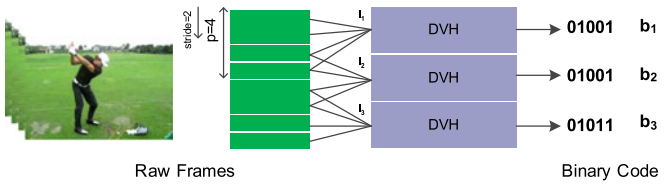

Fig. 3. Extracting binary codes for a single video from DVH. We extract the binary code in a predefined number of consecutive frames, p, and shift to the next set based on a fixed stride value. In this figure, given 8 frames, the DVH fuses $p = 4$ frames with a stride of 2, resulting to 3 final compact binary codes.

TABLE I
DVH IMPLEMENTATIONS IN THE EXPERIMENTS

Early Fusion	Late Fusion	Slow Fusion
fc7–4096	fc7–4096	fc7–4096
pool at p frames	–	pool at p_1 frames
fc8–500	fc8–500	fc8–500
–	pool at p frames	pool at p_2 frames
fc9–200	fc9–200	fc9–200
K	K	K

of binary codes as follows:

$$D_H(\mathbf{X}_i, \mathbf{X}_j) = \frac{1}{g_i g_j} \sum_{u=1}^{g_i} \sum_{v=1}^{g_j} d_H(\mathbf{b}_u, \mathbf{b}_v) \tag{14}$$

where g_i and g_j is the number of frame sets for videos $\mathbf{X}_i$ and $\mathbf{X}_j$. By doing so, we are able to compare the similarity of videos with less computation than using frame to frame comparisons, which are more representative than using single features for each video.

D. Implementation Details

Our deep network composes of a stacked convolutional and pooling layers with parameters obtained from pre-trained models, and connected to a series of fully connected layers such that the top-most layer contains K-dimensional features. The hidden fully connected layers use rectified linear unit (ReLU) as the activation function, while the top-most layer has a hyperbolic tangent activation to ensure centered feature and have balanced $\{-1,1\}$ values. The parameters in the fully connected layers are initialized using the Xavier initialization [62].[3] To be consistent, all models have a fully-connected layers of dimensions $[4096 \rightarrow 500 \rightarrow 200 \rightarrow K]$. To avoid over-fitting, we enable training only in the fully connected layers. Our deep architecture and experiments were implemented under the MatConvNet [63] framework. The learning rate, momentum, and weight decay were set to 0.002, 0.9, and 0.0001, respectively. Table I summarizes the implementations of different DVH methods after the stacked convolutional and pooling layers.

At the training stage, we iteratively passed through all the training videos where we randomly chose video pairs. For each video pair, we randomly chose a set of p successive frames and then packed them into batches to pass into the network. We ensure that the positive and negative pairs for each batch are in an approximate 1:2 ratio. The training procedure converged when the loss does not change within a certain threshold for an epoch. For all experiments, ρ was set to 10 based on the empirical testing to obtain a smooth approximation for the Hinge Loss. We experimented with different values, and found that the results appear to be particularly insensitive to ρ.

[3]We initialize the biases to be 0 and the weights at each layer as $\mathbf{W} = U\left[-\sqrt{\frac{6}{n_{in}+n_{out}}}, \sqrt{\frac{6}{n_{in}+n_{out}}}\right]$ where $\mathbf{W} \in \mathbb{R}^{n_{in} \times n_{out}}$.

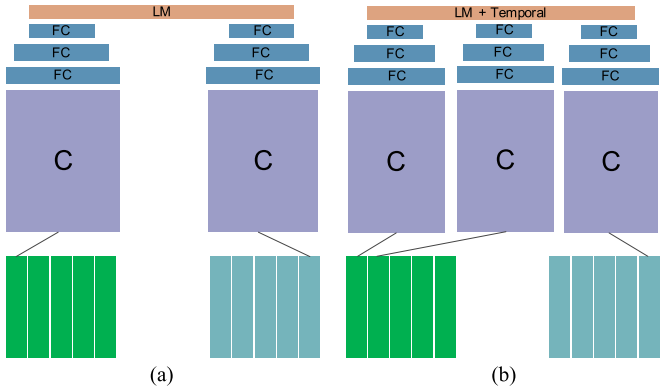

Fig. 4. Deep baseline architectures based on frame-by-frame training for video hashing. The first baseline is similar to image hashing where each frame is a single image. The second baseline adds a temporal criterion during training such that given two frames that are temporally close should have a similar compact feature as much as possible. (a) Single frame. (b) Single frame + temporal.

IV. EXPERIMENTS

To evaluate the effectiveness of our proposed DVH method for scalable video search, we conducted experiments on two video datasets namely the CCV and JHMDB datasets. The details of the experiments and the results are described in the following sections.

A. Datasets and Experimental Settings

Columbia Consumer Video (CCV) dataset [64]: It consists of 9,317 videos with an average duration of 80 seconds extracted from YouTube. The videos were categorized to 20 different categories such as *basketball*, *wedding* and *music performance*. Similar to [26], we sampled frames every 2 seconds and ensured that each video had a minimum of 30 frames. Since most of the categories in this dataset are events, the videos contain large variations among frames making the task very challenging. In our experiments, we randomly selected 20 videos per category for training, 25 videos per category as the query data, and 100 videos per category as the gallery data. This results in 400, 500, and 2000 videos for the training, query, and gallery sets, respectively.

For our deep model, we used the pre-trained VGG-net [53] as our stacked convolution and pooling layers. We used a batch size of 200 and a frame size of $p = 10$. For our Slow Fusion architecture, the first pooling layer fuses the data of $p_1 = 5$ frames with a stride of 2, the second pooling layer fuses the data of the final $p_2 = 3$ frames. For testing samples, we obtained the binary codes for each video at a frame stride of 5.

TABLE II
RESULTS ON THE CCV DATASET IN COMPARISON WITH THE BASELINE DEEP ARCHITECTURE

Method	Hamming ranking (mAP, %)			precision (%) @ N = 100			precision (%) @ r = 2	
	16	32	64	16	32	64	16	32
Single	32.62	34.23	35.02	37.37	38.86	38.50	23.38	5.63
Single+Temporal	33.60	35.60	37.27	38.05	39.74	40.93	30.76	16.44
Video-Level	29.45	30.79	29.19	34.44	35.98	34.05	22.48	11.11
Early Fusion	37.18	40.86	**41.54**	40.11	41.89	**42.41**	36.61	22.80
Late Fusion	**38.54**	**41.08**	41.51	**40.29**	**42.08**	42.23	**37.32**	**23.10**
Slow Fusion	38.27	40.80	41.41	39.95	41.88	42.34	36.55	23.06

TABLE III
RESULTS ON THE JHMDB DATASET IN COMPARISON WITH THE BASELINE DEEP ARCHITECTURE

Method	Hamming ranking (mAP, %)			precision (%) @ N = 10			precision (%) @ r = 2	
	16	32	64	16	32	64	16	32
Single	32.73	33.89	31.74	42.67	43.81	44.05	6.19	0
Single+Temporal	33.19	34.85	35.58	43.81	45.33	45.19	9.05	0
Video-Level	30.07	30.95	32.53	39.10	41.67	41.86	6.58	0
Early Fusion	**35.19**	**37.43**	**37.95**	**46.48**	47.62	48.19	**31.31**	**12.46**
Late Fusion	34.93	36.78	37.53	45.10	**48.05**	48.14	29.67	10.10
Slow Fusion	34.86	36.59	36.63	44.52	47.90	**48.24**	25.25	8.67

Joint-annotated HMDB (JHMDB) dataset [65]: It consists of 928 action videos having 36 to 55 frames per video, which was taken from the HMDB dataset [66] for human motion recognition. The action videos are categorized into 21 human actions such as *brushing hair*, *clapping*, and *climbing*. Although action recognition makes use of flow and RGB information [67], we only used the full body optical flow representation for simplicity. In our experiments, we randomly selected 10 videos per category as training samples, 10 videos per category as query samples, and 20 videos per category as gallery samples. This results in 210, 210, and 420 videos for the training, query, and gallery sets, respectively.

For our deep model, we used the CNN motion network by [68] as our pre-trained stacked convolution and pooling layers. We used a batch size of 50 and a frame size of $p = 10$. For our Slow Fusion architecture, the first pooling layer fuses the data of $p_1 = 5$ frames with a stride of 2, the second pooling layer fuses the data of the final $p_2 = 3$ frames. For testing samples, we obtained the binary codes at a stride of 2.

B. Evaluation Metrics

To measure the performance of our DVH, we used the Hamming ranking and Hamming look-up as evaluation metrics to compare the performance of different methods. For Hamming ranking, the mean Average Precision (mAP) and Precision@N are evaluated. The mAP is defined as the mean of the average precision of the top retrieved samples across all queries, while Precision@N is defined as the percentage of true labels among the top N retrieved samples. For Hamming look-up, the precision when the hamming radius is set as $r = 2$ is evaluated where it measures the precision over all the samples that is within a hamming radius of $r = 2$. At $K = 64$, Hamming look-up precision is not evaluated because it will be impractical for longer bit lengths.

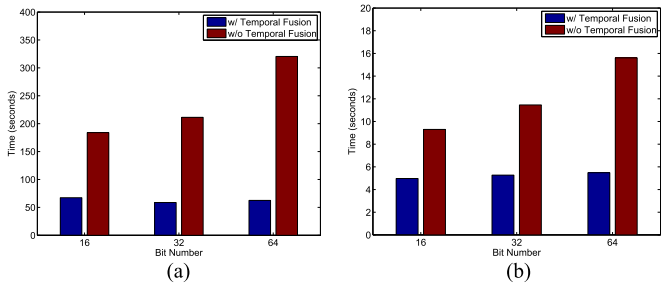

Fig. 5. Retrieval time on the (a) CCV and (b) JHDMB datasets.

C. Experimental Results

Comparison with different deep baselines: We first compared our DVH with three baseline deep architectures which do not use temporal fusion in the fully-connected layers. The baseline methods are described as follows:

Single-frame: In the single-frame model, we considered each frame of the video as a single image with its own label information. Similar to DVH, we used the large-margin criterion for the Siamese network to learn the parameters. However, we only used single frames as the input and do not perform temporal fusion. The cost function is defined as

$$J = f(1 - \delta_{u,v}(\theta - d_{u,v}(s(\mathbf{x}_u), s(\mathbf{x}_v)))). \tag{15}$$

Single-frame + temporal: In the single-fame + temporal model, we exploited the temporal information with the same large-margin criterion so that the frames which are close to each

TABLE IV
RESULTS OF DIFFERENT LEARNING-BASED HASHING METHODS ON THE CCV DATASET

Method	Hamming ranking (mAP, %)			precision (%) @ N = 100			precision (%) @ r = 2	
	16	32	64	16	32	64	16	32
PCAH [28]	20.83	21.45	19.37	25.80	26.50	25.51	3.03	0
PCA-ITQ [6]	22.49	24.13	24.42	27.71	28.99	29.61	13.43	0
AGH [38]	14.91	15.22	11.24	20.52	23.37	20.16	13.43	1.58
KSH [31]	32.43	34.34	35.40	36.27	38.33	38.75	18.27	7.64
CCA-ITQ [6]	36.58	38.18	38.32	39.13	40.41	40.51	16.15	7.17
FastHash [37]	34.72	38.37	38.47	38.83	40.85	41.37	12.73	5.36
DVH	**38.54**	**41.08**	**41.51**	**40.29**	**42.08**	**42.23**	**37.32**	**23.10**

TABLE V
RESULTS OF DIFFERENT LEARNING-BASED HASHING METHODS ON THE JHDMB DATASET

Method	Hamming ranking (mAP, %)			precision (%) @ N = 10			precision (%) @ r = 2	
	16	32	64	16	32	64	16	32
PCAH [28]	16.89	17.64	18.30	27.05	31.31	31.81	0	0
PCA-ITQ [6]	13.80	14.16	14.44	22.52	25.19	27.05	0.58	0
AGH [38]	13.74	14.30	16.90	23.38	26.86	27.57	0.12	0
KSH [31]	27.50	28.32	33.51	39.14	39.05	42.38	0	0
CCA-ITQ [6]	27.20	31.96	31.44	43.48	47.56	47.43	1.51	0
FastHash [37]	31.19	33.72	36.63	42.29	44.33	46.52	0	0
DVH	**35.19**	**37.43**	**37.95**	**46.48**	**47.62**	**48.42**	**31.31**	**12.46**

other are similar as much as possible, defined as below:

$$J = f(1 - \delta_{u_1,v}(\theta - d_{u_1,v}(s(\mathbf{x}_{u_1}), s(\mathbf{x}_v)))) + \nu \|s(\mathbf{x}_{u_1}) - s(\mathbf{x}_{u_2})\|_2^2 \quad (16)$$

where ν is the balancing term, and u_1 and u_2 are two randomly selected image frames from the same video which are apart by a minimum of five frames. In the experiments, we used $\nu = 0.1$. Fig. 4 shows the architectures of the first two baseline methods.

Video-level feature: In this model, we pooled all frame-level features from the single-frame deep model to compute video-level features to evaluate the large margin criterion. By doing so, we obtained a representative global binary vector for each video.

Tables II and III show the performance of different methods on the CCV and JHMDB datasets, respectively. As can be seen, our DVH architectures outperform the other deep learning based baseline architectures. It is interesting to see that the second baseline (Single+Temporal) beats the first baseline (Single), which shows that temporal information is important. The video-level features also yielded competitive representations but often achieved worse performance than the single-frame deep model. The late fusion and early fusion DVH architectures obtain the best performance on the CCV and JHMDB datasets, respectively. For CCV, a late fusion would mean it prefers to exploit high-level global information, while for JHMDB an Early Fusion shows it prefers to exploit local motion information. This is reasonable because JHMDB focuses on action information while CCV is more on events which are generally holistic.

We also examined the retrieval time of different deep baseline architectures, which is shown in Fig. 5. As can be seen, a longer retrieval time is necessary since the deep baseline architectures transform each image frame into a binary code. Differently, our DVH method performs temporal fusion so fewer binary code comparisons are implemented resulting in a faster retrieval time for the whole query-gallery set. This is most obvious on the CCV dataset since more frames are present in a single video, and larger gallery and query videos are used.

Comparison with state-of-the-art learning-based hashing methods: We also compared our DVH method with several popular hashing methods including PCA Hashing [28], PCA-ITQ [6], Anchor Graph Hashing (AGH) [38], Kernel Supervised Hashing (KSH) [31], CCA-ITQ [6], and FastHash [37]. Specifically, PCAH, PCA-ITQ and AGH are unsupervised hashing methods, and KSH, CCA-ITQ, and FastHash exploit the label information of samples to learn discriminative hash codes. The standard implementations of all methods are from the original authors and the default parameters were set based on their respective papers. For consistency, the experiments were carried out with the same selected training, gallery and query sets. For the different hashing methods being compared, we considered each frame as an image and encoded its respective binary code based on the 4096-dimension CNN feature obtained from the fully-connected layer of the pre-trained models used, and defined the hamming distance of two videos as the average of all hamming distances between images from each video.

Tables IV and V show the performance of different hashing methods on the CCV and JHMDB experiments, respectively. We found that the DVH architecture yielded the best performance, where the Late fusion was for the CCV dataset, and the Early fusion was for the JHMDB dataset, respectively. As can be seen, our method consistently outperforms the other

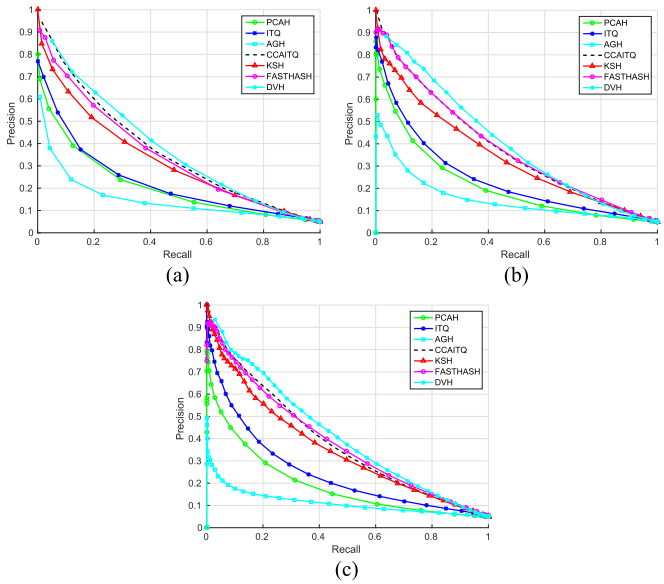

Fig. 6. Precision-recall (PR) curves on the CCV dataset versus varying code lengths. (a) PR curve (16 bits). (b) PR curve (32 bits). (c) PR curve (64 bits).

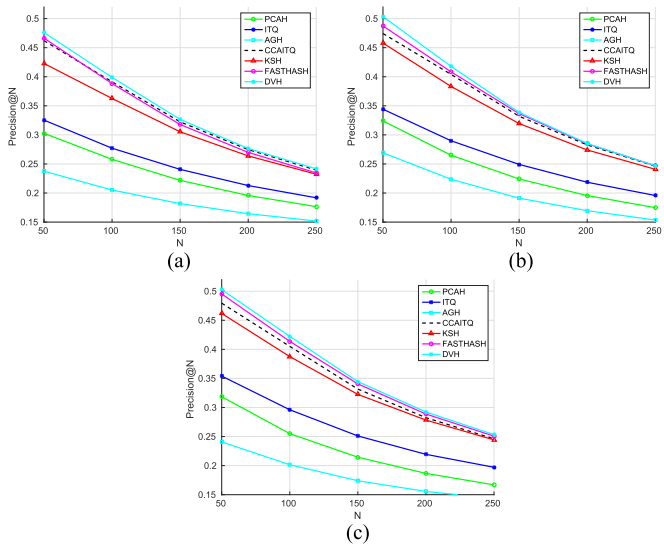

Fig. 7. Precision-N curves on the CCV dataset versus varying code lengths. (a) Prec-N curve (16 bits). (b) Prec-N curve (32 bits). (c) Prec-N curve (64 bits).

existing hashing methods. Most surprising is the hamming look-up precision (HLP) evaluation results which show significant improvement across varying bit lengths. This shows that representing the video in a deep nonlinear binary feature vector gives strong representation for retrieval. Figs. 6–9 show the recall and precision curve, and precision curves vs the retrieval number N on the CCV and JHMDB datasets. We see that our method outperforms the compared methods in most scenarios.

Comparison with different video hashing methods: We compared our method with two video hashing methods as shown in Table VI. For Video Hashing with both Discriminative commonality and Temporal consistency (VHDT) [26], the mAP results were obtained from the original paper. However, their method used SIFT BoW features so it is difficult to directly compare. To approximate, we have applied ITQ-SIFT and found that it

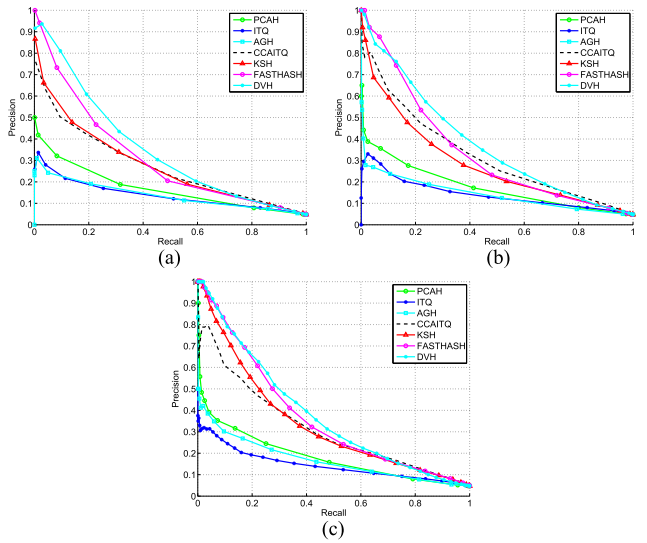

Fig. 8. PR curves on the JHMDB dataset versus varying code lengths. (a) PR curve (16 bits). (b) PR curve (32 bits). (c) PR curve (64 bits).

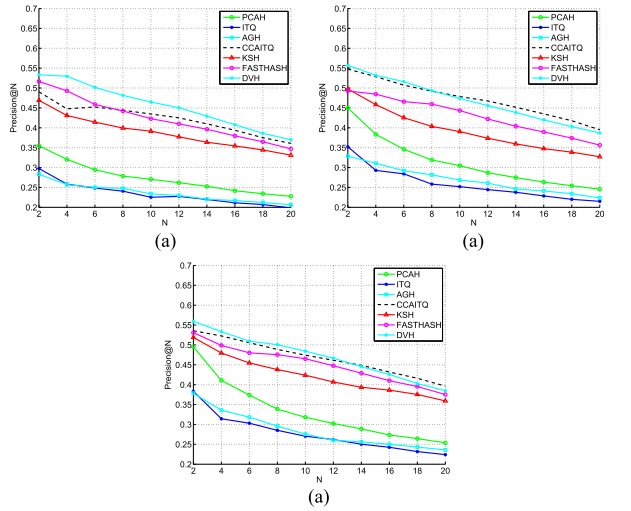

Fig. 9. Precision-N curves on the JHMDB dataset versus varying code lengths. (a) Prec-N curve (16 bits). (b) Prec-N curve (32 bits). (c) Prec-N curve (64 bits).

TABLE VI
HAMMING RANKING (MAP, %) RESULTS ON THE CCV DATASET IN COMPARISON WITH OTHER VIDEO HASHING ALGORITHMS

Method	16 bits	32 bits	64 bits
DVH-CNN	**38.54**	**41.08**	**41.51**
ITQ-CNN	22.49	24.13	24.42
ITQ-SIFT	12.33	14.63	14.52
VHDT-SIFT [26]	11.40	13.00	15.90
CVC-CNN [45]	27.53	32.16	36.14

is comparable with VHDT. Our DVH is much better than ITQ-CNN. This is because VHDT performs linear transformations which may not really capture the nonlinearity of data in videos.

For the Compact Video Coding (CVC) method [45], we used the publicly released code, tuned the parameters to obtain the best possible result and used CNN features to construct the covariance feature for each video. As can be seen, our DVH outperforms CVC at all bit lengths. This is because CVC converted

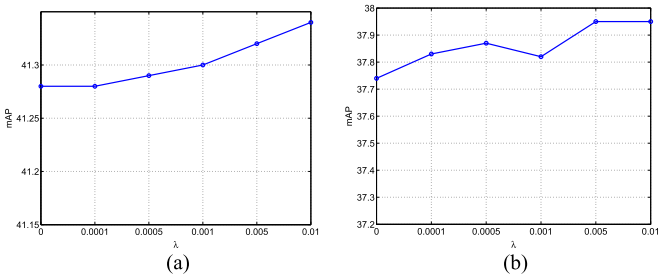

Fig. 10. mAP performance of our DVH method at varying λ for the 64-bit experiment on the (a) CCV and (b) JHMDB datasets, respectively.

TABLE VII
HAMMING RANKING (MAP, %) RESULTS OF OUR DVH ON THE CCV DATASET IN DIFFERENT VALUES OF p AND s

Method	16 bits	32 bits	64 bits
$p = 2, s = 2$	37.81	39.48	40.46
$p = 5, s = 5$	37.81	40.78	41.09
$p = 10, s = 5$	**38.54**	**41.08**	**41.51**
$p = 20, s = 5$	37.25	40.11	41.31

TABLE VIII
HAMMING RANKING (MAP, %) RESULTS OF OUR DVH ON THE JHMDB DATASET IN DIFFERENT VALUES OF p AND s

Method	16 bits	32 bits	64 bits
$p = 2, s = 2$	32.48	32.58	33.55
$p = 5, s = 5$	33.21	35.32	35.30
$p = 10, s = 5$	**35.19**	**37.46**	**37.95**
$p = 20, s = 5$	37.97	35.82	36.43

the whole video into a single feature code which may lead to loss of temporal information, while our DVH method considered the temporal and discriminative information of each video.

Parameter analysis: We also analyzed the varying values of λ during training to see the contribution of the two criterions in the over-all performance of our DVH method. Fig. 10 shows the mAP performance of our DVH method at λ = [0, 0.01, 0.005, 0.001, 0.0005, 0.0001] for the CCV and JHMDB datasets. As expected, we see that the discriminative information makes most of the contribution since even at λ = 0, the performance is very competitive. Nevertheless, minimizing the quantization loss still provides improvement in the over-all performance. However, it is important to see that the quantization loss criterion should not overpower the discriminative criterion. In our experiments, the optimal value for λ is at a range of $[0.005, 0.01]$.

We also conducted experiments on varying values of the frame size p, which is the number of frames as the input in the deep network to obtain a binary code, and the stride s, which is the number of non-overlapped frames. We used the best fusion algorithm for each dataset (CCV-Late, JHMDB-Early). As can be seen in Tables VII and VIII, our method shows a decline in mAP at $p < 10$ probably because frames are very much similar which do not really exploit the temporal information. Similarly, a much higher p may also reduce the performance because it extracts more global video features.

D. Discussion

The above experimental results suggest the following three key observations:

1) Our deep video hashing method achieves very competitive performance compared to other deep baseline architectures which shows that performing temporal fusion during training contributes well to the over-all performance. In addition, retrieval time is also reduced because of the temporal fusion.
2) Our DVH outperforms state-of-the-art image-based hashing methods which shows that the binary codes obtained from our hashing method are strong representations due to the discriminative training we employed. Furthermore, our DVH also outperforms other video hashing methods by a large margin.
3) The large-margin criterion yields the largest contribution in our DVH method. However, the binary quantization term also provides improvements in the over-all performance. For the parameter p, we see that the best performance can be obtained when the parameter of p is set to 10 because it is a good balance of extracting global and local video features.

V. CONCLUSION

In this paper, we have proposed a deep video hashing approach with various frame pooling architectures to learn binary codes for each video in a deep framework such that both temporal and discriminative information are well exploited. Experimental results on two video databases clearly demonstrate that our method achieved better performance with the state-of-the-art hashing methods.

There are two interesting directions for future work:

1) Our DVH method composed of frame-level pooling layers to exploit temporal information. It is interesting to incorporate more complex networks such as recurrent neural networks (RNN) [69], long short term memory (LSTM) [70] and 3D-CNNs [51] to further improve the performance.
2) In this work, we learned our DVH network using supervised information. Hence, it is interesting to learn a deep network using quantization-based [6], [71] criterions, which does not exploit label information.

REFERENCES

[1] S. C. Hoi and M. R. Lyu, "A multimodal and multilevel ranking scheme for large-scale video retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 607–619, Jun. 2008.
[2] C. L. Chou, H. T. Chen, and S. Y. Lee, "Pattern-based near-duplicate video retrieval and localization on web-scale videos," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 382–395, Mar. 2015.
[3] C. Kofler, M. Larson, and A. Hanjalic, "Intent-aware video search result optimization," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1421–1433, Aug. 2014.
[4] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 41, no. 6, pp. 797–819, Nov. 2011.
[5] L. Yu, Z. Huang, J. Cao, and H. T. Shen, "Scalable video event retrieval by visual state binary embedding," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1590–1603, Aug. 2016.

[6] Y. Gong and S. Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in *Proc. 2011 IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 817–824.

[7] K. He, F. Wen, and J. Sun, "K-means hashing: An affinity-preserving quantization method for learning binary compact codes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 2938–2945.

[8] P. Li, M. Wang, J. Cheng, C. Xu, and H. Lu, "Spectral hashing with semantically consistent graph for image indexing," *IEEE Trans. Multimedia*, vol. 15, no. 1, pp. 141–152, Jan. 2013.

[9] W. Liu, C. Mu, S. Kumar, and S.-F. Chang, "Discrete graph hashing," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 3419–brk?>3427.

[10] W. Zhou *et al.*, "Towards codebook-free: Scalable cascaded hashing for mobile image search," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 601–611, Apr. 2014.

[11] Y. Li, R. Wang, S. Shan, and X. Chen, "Hierarchical hybrid statistic based video binary code and its application to face retrieval in tv-series," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recog.*, May 2015, pp. 1–8.

[12] X. Liu, L. Huang, C. Deng, B. Lang, and D. Tao, "Query-adaptive hash code ranking for large-scale multi-view visual search," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4514–4524, Oct. 2016.

[13] X. Liu *et al.*, "Multilinear hyperplane hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 5119–5127.

[14] J. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen, "A survey on learning to hash," *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1606.00185 2016.

[15] J. Masci, M. M. Bronstein, A. M. Bronstein, and J. Schmidhuber, "Multimodal similarity-preserving hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 824–830, Apr. 2014.

[16] J. Zhou, G. Ding, and Y. Guo, "Latent semantic sparse hashing for cross-modal similarity search," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2014, pp. 415–424.

[17] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 3864–3872.

[18] F. Wu *et al.*, "Sparse multi-modal hashing," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 427–439, Feb. 2014.

[19] X. Xu, F. Shen, Y. Yang, and H. T. Shen, "Discriminant cross-modal hashing," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2016, pp. 305–308.

[20] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2008, pp. 1753–1760.

[21] A. Gionis *et al.*, "Similarity search in high dimensions via hashing," in *Proc. 25th Int. Conf. Very Large Data Bases*, 1999, pp. 518–529.

[22] R. Ji *et al.*, "Learning to distribute vocabulary indexing for scalable visual search," *IEEE Trans. Multimedia*, vol. 15, no. 1, pp. 153–166, Jan. 2013.

[23] J. Song, Y. Yang, Z. Huang, H. T. Shen, and J. Luo, "Effective multiple feature hashing for large-scale near-duplicate video retrieval," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1997–2008, Dec. 2013.

[24] M. Douze, H. Jégou, and C. Schmid, "An image-based approach to video copy detection with spatio-temporal post-filtering," *IEEE Trans. Multimedia*, vol. 12, no. 4, pp. 257–266, Jun. 2010.

[25] B. Coskun, B. Sankur, and N. Memon, "Spatio–Temporal transform based video hashing," *IEEE Trans. Multimedia*, vol. 8, no. 6, pp. 1190–1208, Dec. 2006.

[26] G. Ye, D. Liu, J. Wang, and S.-F. Chang, "Large-scale video hashing via structure learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2272–2279.

[27] L. Cao, Z. Li, Y. Mu, and S.-F. Chang, "Submodular video hashing: A unified framework towards video pooling and indexing," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 299–308.

[28] J. Wang, S. Kumar, and S.-F. Chang, "Semi-supervised hashing for scalable image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 3424–3431.

[29] G. Irie, Z. Li, X.-M. Wu, and S.-F. Chang, "Locally linear hashing for extracting non-linear manifolds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 2115–2122.

[30] F. Shen *et al.*, "Hashing on nonlinear manifolds," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1839–1851, Jun. 2015.

[31] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang, "Supervised hashing with kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2074–2081.

[32] Y. Mu, G. Hua, W. Fan, and S.-F. Chang, "Hash-SVM: Scalable kernel machines for large-scale visual classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 979–986.

[33] L.-Y. Duan, J. Lin, Z. Wang, T. Huang, and W. Gao, "Weighted component hashing of binary aggregated descriptors for fast visual search," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 828–842, Jun. 2015.

[34] X. Liu, Y. Mu, D. Zhang, B. Lang, and X. Li, "Large-scale unsupervised hashing with shared structure learning," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1811–1822, Sep. 2015.

[35] F. Shen, C. Shen, Q. Shi, A. Van Den Hengel, and Z. Tang, "Inductive hashing on manifolds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 1562–1569.

[36] Y.-G. Jiang, J. Wang, X. Xue, and S.-F. Chang, "Query-adaptive image search with hash codes," *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 442–453, Feb. 2013.

[37] G. Lin, C. Shen, Q. Shi, A. Hengel, and D. Suter, "Fast supervised hashing with decision trees for high-dimensional data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1963–1970.

[38] W. Liu, J. Wang, S. Kumar, and S.-F. Chang, "Hashing with graphs," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 1–8.

[39] V. Erin Liong, J. Lu, G. Wang, P. Moulin, and J. Zhou, "Deep hashing for compact binary codes learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 2475–2483.

[40] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4766–4779, Dec. 2015.

[41] H. Zhu, M. Long, J. Wang, and Y. Cao, "Deep hashing network for efficient similarity retrieval," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 2415–2421.

[42] F. Zhao, Y. Huang, L. Wang, and T. Tan, "Deep semantic ranking based hashing for multi-label image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 1556–1564.

[43] Y. Cao, M. Long, J. Wang, H. Zhu, and Q. Wen, "Deep quantization network for efficient image retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 3457–3463.

[44] J. Wang, J. Sun, J. Liu, X. Nie, and H. Yan, "A visual saliency based video hashing algorithm," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep.-Oct. 2012, pp. 645–648.

[45] Y. Li, R. Wang, Z. Cui, S. Shan, and X. Chen, "Compact video code and its application to robust face retrieval in tv-series," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–12.

[46] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1701–1708.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[48] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[49] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 580–587.

[50] X. Zeng, W. Ouyang, M. Wang, and X. Wang, "Deep learning of scene-specific classifier for pedestrian detection," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 472–487.

[51] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[52] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *Proc. 2nd Int. Conf. Human Behavior Understanding*, 2011, pp. 29–39.

[53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[54] X. Peng and C. Schmid, "Multi-region two-stream R-CNN for action detection," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 744–759.

[55] A. Karpathy *et al.*, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1725–1732.

[56] J. Yue-Hei Ng *et al.*, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 4694–4702.

[57] H. Ye *et al.*, "Evaluating two-stream CNN for video classification," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2015, pp. 435–442.

[58] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative CNN video representation for event detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 1798–1807.

[59] J. Shao, C.-C. Loy, K. Kang, and X. Wang, "Slicing convolutional neural network for crowd video understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 5620–5628.
[60] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, pp. 539–546.
[61] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, 2002.
[62] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
[63] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 689–692.
[64] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui, "Consumer video understanding: A benchmark database and an evaluation of human and machine performance," in *Proc. ACM Int. Conf. Multimedia*, 2011, pp. 29–37.
[65] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. Black, "Towards understanding action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3192–3199.
[66] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.
[67] G. Chéron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN features for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3218–3226.
[68] G. Gkioxari and J. Malik, "Finding action tubes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 759–768.
[69] D. Rumelhart, G. Hinton, and R. Williams, "Learning sequential structure in simple recurrent networks," *Parallel Distributed Processing: Experiments in the Microstructure of Cognition*. Cambridge, MA, USA: MIT Press, 1986.
[70] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[71] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 117–128, Jan. 2011.

Venice Erin Liong received the B.S. degree from the University of the Philippines Diliman, Quezon City, Philippines, in 2010, the M.S. degree from the Korea Advanced Institute of Science and Technology, Daejeon City, South Korea, in 2013, and is currently working toward the Ph.D. degree at the Rapid-Rich Object Search (ROSE) Laboratory, Interdisciplinary Graduate School, Nanyang Technological University, Singapore.

Her research interests include computer vision and pattern recognition.

Jiwen Lu (S'10–M'11–SM'15) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from the Nanyang Technological University, Singapore, in 2012.

He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. From March 2011 to November 2015, he was a Research Scientist with the Advanced Digital Sciences Center, Singapore. He has authored or coauthored more than 130 scientific, 35 of them IEEE Transactions papers. His current research interests include computer vision, pattern recognition, and machine learning.

Prof. Lu is a Member of the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society. He serves/has served as an Associate Editor of *Pattern Recognition Letters, Neurocomputing*, and IEEE ACCESS, a Managing Guest Editor of *Pattern Recognition and Image and Vision Computing*, and a Guest Editor of *Computer Vision and Image Understanding*. He is/was a Workshop Chair/Special Session Chair/Area Chair for more than ten international conferences. He was the recipient of the National 1000 Young Talents Plan Program in 2015.

Yap-Peng Tan (S'95–M'98–SM'04) received the B.S. degree from National Taiwan University, Taipei, Taiwan, in 1993, and the M.A. and Ph.D. degrees from Princeton University, Princeton, NJ, USA, in 1995 and 1997, respectively, all in electrical engineering.

From 1997 to 1999, he was with Intel Corporation, Chandler, AZ, USA, and Sharp Laboratories of America, Camas, WA, USA. In November 1999, he joined the Nanyang Technological University of Singapore, Singapore, where he is currently an Associate Professor and Associate Chair (Academic) of the School of Electrical and Electronic Engineering. He is the principal inventor or co-inventor on 15 U.S. patents in the areas of image and video processing. His current research interests include image and video processing, content-based multimedia analysis, computer vision, pattern recognition, and data analytics.

Prof. Tan served as the Chair of the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society from 2012 to 2014, a Member of the Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society from 2009 to 2013, Voting Member of the IEEE International Conference on Multimedia & Expo Steering Committee from 2011 to 2012, and Chairman of the IEEE Signal Processing Singapore Chapter from 2009 to 2010. He has also served as an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS (since 2016), the IEEE TRANSACTIONS ON MULTIMEDIA (since 2014), and IEEE ACCESS (since 2013), an Editorial Board Member of the EURASIP *Journal on Advances in Signal Processing* and the EURASIP *Journal on Image and Video Processing*, a Guest Editor for special issues of several journals including the IEEE TRANSACTIONS ON MULTIMEDIA, and a Member of the Multimedia Systems and Applications Technical Committee and Visual Signal Processing and Communications Technical Committee (VSPC TC) of the IEEE Circuits and Systems Society. He is the Tutorial Co-Chair of the 2016 IEEE International Conference on Multimedia and Expo (ICME 2016) and Technical Program Co-Chair of the 2019 IEEE International Conference on Image Processing (ICIP 2019), and was the Finance Chair of the 2004 IEEE International Conference on Image Processing (ICIP 2004), General Co-Chair of the 2010 IEEE International Conference on Multimedia and Expo (ICME 2010), Technical Program Co-Chair of the 2015 IEEE International Conference on Multimedia and Expo (ICME 2015), and General Co-Chair of the 2015 IEEE International Conference on Visual Communications and Image Processing (VCIP 2015).

Jie Zhou (M'01–SM'04) received the B.S. and M.S. degrees in mathematics from Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 1995.

From then to 1997, he served as a Postdoctoral Fellow with the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a Full Professor with the Department of Automation, Tsinghua University. In recent years, he has authored or coauthored more than 100 papers in peer-reviewed journals and conferences. Among them, more than 40 papers have been published in top journals and conferences such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and CVPR. His research interests include computer vision, pattern recognition, and image processing.

Prof. Zhou is an Associate Editor for the *International Journal of Robotics and Automation* and two other journals. He was the recipient of the National Outstanding Youth Foundation of China Award.