Face Video Retrieval with Image Query via Hashing across Euclidean Space and Riemannian Manifold

Yan Li[1,2], Ruiping Wang[1], Zhiwu Huang[1,2], Shiguang Shan[1], Xilin Chen[1,3]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China.
[2]University of Chinese Academy of Sciences, Beijing, 100049, China.
[3]Department of Computer Science and Engineering, University of Oulu, Oulu 90570, Finland.

Retrieving videos of a specific person given his/her face image as query becomes more and more appealing for applications like smart movie fast-forwards and suspect searching. It also forms an interesting but challenging computer vision task, as the visual data to match, i.e., still image and video clip are usually represented quite differently. Typically, face image is represented as point (i.e., vector) in Euclidean space, while video clip is seemingly modeled as a point (e.g., covariance matrix) on some particular Riemannian manifold in the light of its recent promising success. It thus incurs a new hashing-based retrieval problem of matching two heterogeneous representations, respectively in Euclidean space and Riemannian manifold.

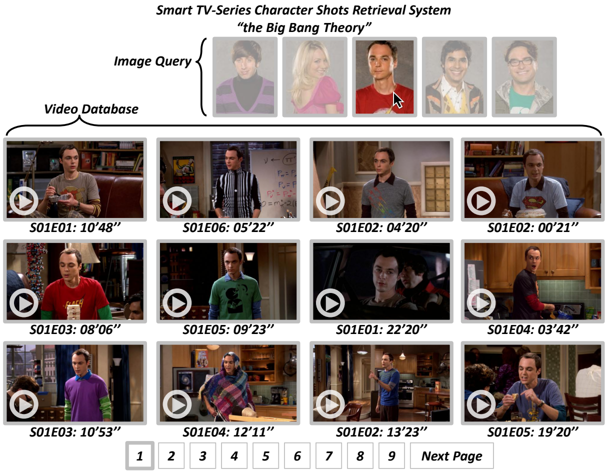

Figure 1: A conceptual illustration of TV-Series (*the Big Bang Theory*) character shots retrieval, where the query is an image of one specific character (*Sheldon Cooper*), and all the shots containing him/her are retrieved and ranked according to their similarities to the query image.

This work makes the first attempt to embed the two heterogeneous spaces into a common discriminant Hamming space. Specifically, we propose Hashing across Euclidean space and Riemannian manifold (HER) by deriving a unified framework to firstly embed the two spaces into corresponding reproducing kernel Hilbert spaces (Fig. 2), and then iteratively optimize the intra- (Eqn. (2) and Eqn. (3)) and inter-space (Eqn. (4)) Hamming distances in a max-margin framework to learn the hash functions for the two spaces.

$$
\begin{aligned}
\min_{W_e,W_r,\xi_e,\xi_r,B_e,B_r} \quad & \lambda_1 E_e + \lambda_2 E_r + \lambda_3 E_{er} \\
& + \gamma_1 \sum_{k\in\{1:K\}} \left\| w_e^k \right\|^2 + C_1 \sum_{\substack{k\in\{1:K\}\\ i\in\{1:N\}}} \xi_e^{ki} \\
& + \gamma_2 \sum_{k\in\{1:K\}} \left\| w_r^k \right\|^2 + C_2 \sum_{\substack{k\in\{1:K\}\\ i\in\{1:N\}}} \xi_r^{ki}
\end{aligned}
\tag{1}
$$

$$
\begin{aligned}
s.t. B_e^{ki} &= sgn({w_e^k}^T \varphi(x_i)), \forall k \in \{1:K\}, i \in \{1:N\} \\
B_r^{ki} &= sgn({w_r^k}^T \eta(\mathcal{Y}_i)), \forall k \in \{1:K\}, i \in \{1:N\} \\
B_r^{ki}({w_e^k}^T \varphi(x_i)) &\geq 1 - \xi_e^{ki}, \xi_e^{ki} > 0, \forall k \in \{1:K\}, i \in \{1:N\} \\
B_e^{ki}({w_r^k}^T \eta(\mathcal{Y}_i)) &\geq 1 - \xi_r^{ki}, \xi_r^{ki} > 0, \forall k \in \{1:K\}, i \in \{1:N\}.
\end{aligned}
$$

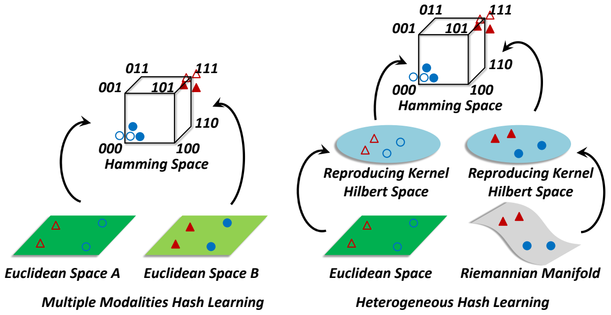

Figure 2: The difference between traditional multiple modalities hash learning methods (the left figure) and our heterogeneous hash learning method (the right figure), where different shapes (i.e., triangles and circles) denote categories.

$$
E_e = \sum_{c\in\{1:C\}} \sum_{m,n\in c} d(B_e^m, B_e^n) - \lambda_e \sum_{\substack{c_1\in\{1:C\}\\ p\in c_1}} \sum_{\substack{c_2\in\{1:C\}\\ c_1\neq c_2, q\in c_2}} d(B_e^p, B_e^q) \tag{2}
$$

$$
E_r = \sum_{c\in\{1:C\}} \sum_{m,n\in c} d(B_r^m, B_r^n) - \lambda_r \sum_{\substack{c_1\in\{1:C\}\\ p\in c_1}} \sum_{\substack{c_2\in\{1:C\}\\ c_1\neq c_2, q\in c_2}} d(B_r^p, B_r^q) \tag{3}
$$

$$
E_{er} = \sum_{c\in\{1:C\}} \sum_{m,n\in c} d(B_e^m, B_r^n) - \lambda_{er} \sum_{\substack{c_1\in\{1:C\}\\ p\in c_1}} \sum_{\substack{c_2\in\{1:C\}\\ c_1\neq c_2, q\in c_2}} d(B_e^p, B_r^q) \tag{4}
$$

To evaluate HER, we conduct video face retrieval experiments on two hot American TV-Series, i.e., *the Big Bang Theory* and *Buffy the Vampire Slayer*. Extensive experimental results (Table 1) demonstrate the superiority of HER over the state-of-the-art competitive hash learning methods, and such superiority mainly benefits from three points: 1) the integration of intra- and inter-space discriminability constraints (i.e., E_e, E_r, and E_{er}) via an iterative optimization based on Hamming distance; 2) the two-step architecture, i.e., Euclidean space (Riemannian manifold) to RKHS and then to common Hamming space, involves nonlinear maps from the original spaces into high dimensional Hilbert spaces, which would yield much richer representations of the original data distributions; 3) the max-margin strategy accomplished by SVM further ensures the stability and generalizability of the learned hash functions, which is a crucial element for practical retrieval system.

Table 1: Comparison with the state-of-the-art single modality and multiple modalities hash learning methods with mAP on *the Big Bang Theory*. *K* means the length of hash code.

Hashing Method	*the Big Bang Theory*			
	$K = 16$	$K = 32$	$K = 64$	$K = 128$
LSH [Indyk & Motwani, STC'98]	0.2086	0.2092	0.1963	0.1994
SH [Weiss, NIPS'08]	0.2652	0.2665	0.2623	0.2673
ITQ [Gong, CVPR'11]	0.3025	0.2989	0.3029	0.3060
SSH [Wang, CVPR'10]	0.2855	0.2662	0.2584	0.2586
DBC [Rastegari, ECCV'12]	0.4495	0.4235	0.4005	0.3867
KSH [Liu, CVPR'12]	0.4366	0.4454	0.4567	0.4604
SITQ [Gong, CVPR'11]	0.3909	0.4298	0.4576	0.4799
CMSSH [Bronstein, CVPR'10]	0.2047	0.2143	0.2024	0.2478
CVH [Kumar & Udupa, IJCAI'11]	0.2110	0.2092	0.2231	0.2407
PLMH [Zhai, IJCAI'13]	0.2447	0.2461	0.2487	0.2608
PDH [Rastegari, ICML'13]	0.2949	0.2903	0.3095	0.2916
MLBE [Zhen & Yeung, KDD'12]	0.2600	0.2648	0.3917	0.3858
MM-NN [Masci, PAMI'13]	0.3955	0.4664	0.5124	0.4922
HER	**0.5049**	**0.5227**	**0.5490**	**0.5539**

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

The matlab implementation of HER can be downloaded from `http://vipl.ict.ac.cn/resources/codes`.