Compact Video Code and Its Application to Robust Face Retrieval in TV-Series

Yan Li
yan.li@vipl.ict.ac.cn
Ruiping Wang
wangruiping@ict.ac.cn
Zhen Cui
zhen.cui@vipl.ict.ac.cn
Shiguang Shan
sgshan@ict.ac.cn
Xilin Chen
xlchen@ict.ac.cn

Key Lab of Intelligent Information Processing, Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China

Problem: We address the problem of video face retrieval in TV-Series which searches video clips based on the presence of specific character, given one video clip of his/hers, see Figure 1. This is tremendously challenging because on one hand, faces in TV-Series are captured in largely uncontrolled conditions with complex appearance variations, and on the other hand retrieval task typically needs efficient representation with low time and space complexity.

Our Method: To solve this problem, we propose a compact and discriminative representation for the huge body of video data, named Compact Video Code (CVC). Our method first models the video clip by its sample (i.e., frame) covariance matrix to capture the video data variations in a statistical manner. Let $F = [f_1, f_2, ..., f_n]$ be the data matrix of a video clip with n frames, where $f_i \in \mathbb{R}^d$ denotes the i^{th} frame with d-dimensional feature. We represent the video clip with the $d \times d$ sample covariance matrix:

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(f_i - \bar{f})(f_i - \bar{f})^T, \tag{1}$$

where $\bar{f}$ is the mean of all frames in the video clip. It is well known that the nonsingular covariance matrices do not lie in a Euclidean space but on a Riemannian manifold $\mathcal{M}$. However, it is not trivial to learn a binary code encoder on the manifold since typical code learning methods are devoted to operating in Euclidean space. So here we utilize the Log-Euclidean Distance (LED) to bridge the gap between Riemannian manifold and Euclidean space as in [2]:

$$d_{LED}(C_1, C_2) = \|\log(C_1) - \log(C_2)\|_F. \tag{2}$$

To incorporate discriminative information and obtain more compact video signature, the high-dimensional covariance matrix is further encoded as a much lower-dimensional binary vector, which finally yields the proposed CVC. Specifically, each bit of the code, i.e., each dimension of the binary vector, is produced via supervised learning in a max margin framework [1], which aims to make a balance between the ***discriminability*** and ***stability*** of the code.

Discriminability: We characterize the discriminability into two parts: within class compactness (S_W) and between class separability (S_B).

$$S_W = \sum_{c\in\{1:M\}}\sum_{m,n\in c} dis(b_m, b_n), \tag{3}$$

$$S_B = \sum_{\substack{c_1\in\{1:M\}\\ p\in c_1}}\sum_{\substack{c_2\in\{1:M\}\\ c_1\neq c_2, q\in c_2}} dis(b_p, b_q), \tag{4}$$

where M is the total number of training classes, $dis(\cdot)$ is the distance measurement of binary codes in Hamming space, $b \in \{-1,1\}^{N\times K}$ denotes the binary codes of training instances, and N and K denote the total number of training instances and the length of binary code respectively. Thus, to implement a strong discrimination, we should minimize the following energy function E_{disc}.

$$E_{disc} = S_W - \lambda_1 S_B. \tag{5}$$

Stability: To make better stability, we build the K hyperplanes by using SVM, and each generates one bit of the binary code. Concretely, we

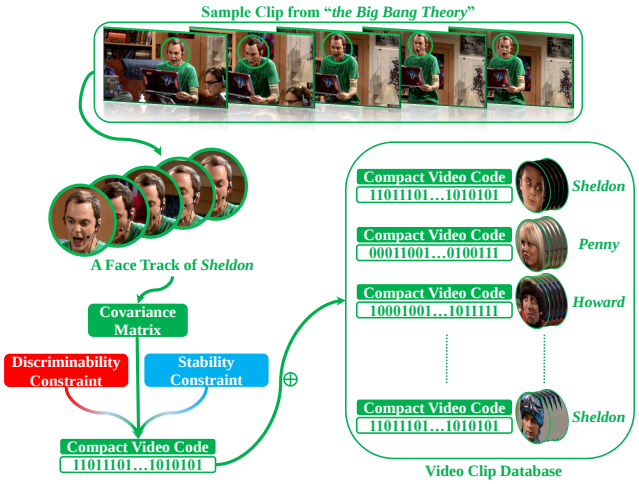

Figure 1: Illustration of the proposed method. Given a video clip of one character as query, we extract the proposed Compact Video Code (CVC) to represent it and use Hamming distance to retrieve video clips containing the specific character in database, which are also encoded in the form of CVC representations.

denote the k^{th} hyperplane by ω^k $(k = 1, \cdots, K)$, and the energy function can be formulated as follow.

$$E_{stab} = \frac{1}{2}\sum_{k\in\{1:K\}} {\omega^k}^T \omega^k + \lambda_2 \sum_{\substack{k\in\{1:K\}\\ i\in\{1:N\}}} \max(1 - b_i^k({\omega^k}^T x_i), 0), \tag{6}$$

where x_i denotes the input feature, b_i^k indicates in which side of the k^{th} hyperplane the i^{th} training instance lies, and λ_2 balances the empirical training error and the hyperplane margin.

After the above analysis, we can reach the final objective function by combining Eqn. (5) and Eqn. (6) to simultaneously consider the discriminability and stability of the target binary code:

$$\min_{b,\omega} \quad E_{disc} + E_{stab}. \tag{7}$$

Since the objective function is non-convex, in practice we independently optimize each individual component to iteratively update b and ω, where an efficient subgradient descend method proposed in [1] with compute-complexity $O(NK)$ was utilized to optimize b.

Experiment: Face retrieval experiments on two challenging TV-Series video databases have demonstrated the competitiveness of the proposed CVC over state-of-the-art retrieval methods. In addition, as a general video matching algorithm, our method is also evaluated in traditional video face recognition task on a standard Internet database, i.e., *YouTube Celebrities*, showing its quite promising performance by using an extremely compact code with only 128 bits.

[1] Mohammad Rastegari, Ali Farhadi, and David Forsyth. Attribute discovery via predictable discriminative binary codes. In *ECCV*, pages 876–889. Springer, 2012.

[2] Ruiping Wang, Huimin Guo, Larry S. Davis, and Qionghai Dai. Covariance discriminative learning: a natural and efficient approach to image set classification. In *CVPR*, pages 2496–2503. IEEE, 2012.