Bilinear Heterogeneous Information Machine for RGB-D Action Recognition

Yu Kong[1], Yun Fu[1,2]
[1]Department of Electrical and Computer Engineering, [2]College of Computer and Information Science. Northeastern University, Boston, MA, USA.

Action recognition from RGB-D cameras has been receiving increasing interests in the computer vision community due to the recent advance of easy-to-use and low-cost depth sensors such as Kinect sensors [3]. Previous studies [1, 2, 3, 4] show that effective usage of 3D structural information provided by RGB-D cameras facilitates recognition tasks as it simplifies intra-class motion variations and removes cluttered background noise. Despite its effectiveness, these methods are only applicable when depth data are available. Methods developed in [1, 2, 5, 6] would fail if depth data are unavailable. In addition, depth data are noisy with discontinuities, which hinders the application of feature extraction methods. Moreover, spatiotemporal correlation information of body part movements would collapse if a vector form representation is used.

In this paper, we propose a novel bilinear heterogeneous information machine (BHIM) for action recognition from RGB-D sequences. BHIM learns cross-modal features that effectively capture heterogeneous RGB and depth information. RGB and depth data are treated as two modalities in this work, and are represented in a matrix format, which naturally encodes spatiotemporal structural relationships. We project the original features of the two modalities onto a shared space, and learn cross-modal features shared between them for classification in order to effectively capture cross-modal knowledge. The learned cross-modal features inherit the characteristics of both RGB and depth data that capture motion, 3D structural, and spatiotemporal relationship information. Moreover, the features are "filtered" for noise removal in the projection procedure. The recognition problem is formulated in a low-rank bilinear framework, particularly designed for the feature representation in a matrix form.

Denote N RGB-D action videos for training purpose by $\{X_i, y_i\}_{i=1}^N$, where $X_i = \{X_i^{[v]}, X_i^{[z]}\} \in \mathcal{X}$ contains a RGB visual feature matrix $X_i^{[v]} \in \mathcal{X}_v$ and a depth feature matrix $X_i^{[z]} \in \mathcal{X}_z$ extracted from RGB-D data, and $y_i \in \mathcal{Y}$ is the corresponding action label. Suppose we are given M ($M = 2$ in this work) types of modalities $X_i^{[m]}|_{m=1}^M$. In this paper, we are interested in a binary bilinear discriminant function $F(X_i, y|W) = \mathrm{Tr}(W^{\mathrm{T}} X_i) = \sum_{m=1}^M \mathrm{Tr}(W_f^{[m]} W_w^{\mathrm{T}} X_i^{[m]})$, which is a family of bilinear functions parameterized by model weight matrix W. Here, parameter matrix $W_f^{[m]} \in \mathcal{R}^{n_f \times d}$ ($m = 1, \cdots, M$) projects the m modality data, $X^{[m]}$, into a learned shared space, and parameter matrix $W_w \in \mathcal{R}^{n_{xyt} \times d}$ is applied to classify the projected data regardless of modalities (see Figure 1).

We train the bilinear model in a max-margin framework. Based on the empirical risk minimization principle, we formulate our learning problem as

$$\min_{W_w, W_f^{[v]}, W_f^{[z]}} \phi(W_f^{[v]}, W_f^{[z]}) + \lambda \cdot r(W_w, W_f^{[v]}, W_f^{[z]}) + C \cdot l(W_w, W_f^{[v]}, W_f^{[z]}), \quad (1)$$

where $\phi(\cdot)$ is a regularizer term for reducing noise in the projected data, $r(\cdot)$ is an additional regularizer term related to the margin of the bilinear model, and $l(\cdot)$ computes training loss. λ and C are trade-off parameters balancing the importance of the corresponding terms.

Regularizer $\phi(W_f^{[v]}, W_f^{[z]})$ attempts to summarize and compress the original two-modality data. Since the raw RGB and depth data may not be in the same space, we use this term to compress the data and discover the shared knowledge between two modalities. We define this term as

$$\phi(W_f^{[v]}, W_f^{[z]}) = I(X^{[v]}, O) + I(X^{[z]}, O), \quad (2)$$

where $X^{[m]} = \{X_i^{[m]}\}_{i=1}^N$ ($m = v$ or $m = z$) represents a set of all training samples in the m modality, $O = \frac{1}{2}(X^{[v]} W_f^{[v]} + X^{[z]} W_f^{[z]}) \in \mathcal{O}$ is the learned low-dimensional cross-modal features in the shared space, and $I(\cdot,\cdot)$ computes mutual information.

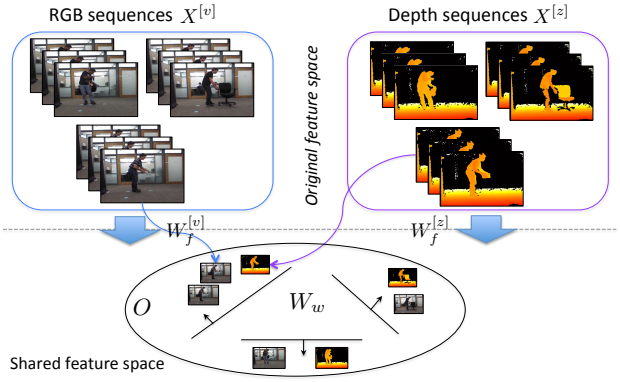

Figure 1: Graphical illustration of the proposed BHIM model.

Regularizer $r(W_w, W_f^{[v]}, W_f^{[z]})$ is used to measure the margin of the bilinear classifier. Minimizing $r(W_w, W_f^{[v]}, W_f^{[z]})$ is equivalent to maximizing the margin of the bilinear model, thereby capturing discriminative information. We define this term as

$$r(W_w, W_f^{[v]}, W_f^{[z]}) = \frac{1}{2}\mathrm{Tr}(W_w W_f^{[v]\mathrm{T}} W_f^{[v]} W_w^{\mathrm{T}}) + \frac{1}{2}\mathrm{Tr}(W_w W_f^{[z]\mathrm{T}} W_f^{[z]} W_w^{\mathrm{T}}). \quad (3)$$

Loss function $l(W_w, W_f^{[v]}, W_f^{[z]})$ computes training loss given the learned model parameter matrices. We consider a binary classifier in this work, and define a hinge loss function for each modality:

$$\begin{aligned} l(W_w, W_f^{[v]}, W_f^{[z]}) = \sum_i \Big[& \max(0, 1 - y_i \mathrm{Tr}(W_f^{[v]} W_w^{\mathrm{T}} X_i^{[v]}) \\ & + \max(0, 1 - y_i \mathrm{Tr}(W_f^{[z]} W_w^{\mathrm{T}} X_i^{[z]}) \Big]. \end{aligned} \quad (4)$$

Plugging Eq. (2), Eq. (3), and Eq. (4) into Eq. (1), the optimal parameter matrices $W_f^{[v]}$, $W_f^{[z]}$ and W_w can be learned by

$$\begin{aligned} \min_{W_w, W_f^{[v]}, W_f^{[z]}} & \sum_m \Big[I(X^{[m]}, O) + \frac{1}{2}\lambda \cdot \mathrm{Tr}(W_w W_f^{[m]\mathrm{T}} W_f^{[m]} W_w^{\mathrm{T}}) + C \cdot \sum_i \xi_i^{[m]} \Big], \\ \text{s.t.} \quad & y_i \mathrm{Tr}(W_f^{[m]} W_w^{\mathrm{T}} X_i^{[m]}) \geqslant 1 - \xi_i^{[m]}, \quad \xi_i^{[m]} \geqslant 0, \quad \forall i, \forall m, \end{aligned} \quad (5)$$

where $\xi_i^{[m]}$ is a slack variable for the m modality in the i-th RGB-D video. This constrained optimization problem can be solved by a coordinate descent algorithm that solves for one set of parameter matrices at each step with the others fixed.

[1] Simon Hadfield and Richard Bowden. Hollywood 3d: Recognizing actions in 3d natural scenes. In *CVPR*, 2013.

[2] Omar Oreifej and Zicheng Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. 2013.

[3] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, and Andrew Blake. Efficient human pose estimation from single depth images. *PAMI*, 2013.

[4] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012.

[5] Lu Xia and J.K. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *CVPR*, 2013.

[6] Xiaodong Yang and YingLi Tian. Super normal vector for activity recognition using depth sequences. In *CVPR*, 2014.

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