# Feature Combination beyond Basic Arithmetics

Hao Fu[1]
hxf@cs.nott.ac.uk
Guoping Qiu[1]
qiu@cs.nott.ac.uk
Hangen He[2]
hehangen@yahoo.com

[1] School of Computer Science
University of Nottingham
Nottingham, UK

[2] College of Mechatronic Engineering and Automation, National University of Defense Technology, Changsha, Hunan, P.R.China, 410073

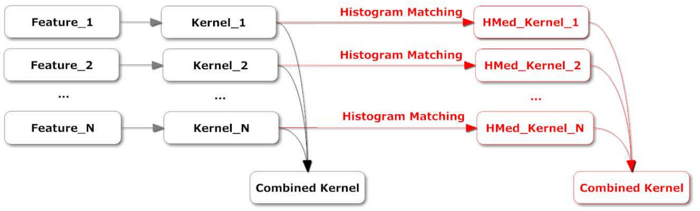

Figure 1: Typical feature combination methods directly combine the kernels through one of the methods in (1) or (2). In this paper, we proposed to add a histogram matching module before these kernels are combined.

Kernel-based feature combination techniques such as Multiple Kernel Learning use arithmetical operations to linearly combine different kernels. The seminal work of Multiple Kernel Learning dates back to [1], where the authors proposed an efficient algorithm to solve this optimization problem. After MKL was proposed, many variants of it have been proposed [4, 7], and have been quickly adopted to deal with various computer vision problems [5, 6]. However, in some scenario, simply average different kinds of kernels may even outperform the sophisticated MKL methods [2].

Mathematically speaking, the baseline average kernel can be written as:

$$k^*(x,x') = \frac{1}{F}\sum_{m=1}^{F} k_m(x,x') \tag{1}$$

In the case of MKL, the combined kernel $k^*$ is a linear combination of different kernels weighted by a set of parameters $\{\beta_m\}$ to be learned by the MKL algorithms:

$$k^*(x,x') = \sum_{m=1}^{F} \beta_m k_m(x,x') \tag{2}$$

In this paper, we make an important observation of the distribution of different kernels that are routinely used in the literature. We discovered that the histograms of different kernels are usually quite different from each other (see Fig.2 for example). Some histograms may be narrow and occupy only a short range, while others may span a wide range; some histograms may look like a gaussian distribution, while others may look like an exponential distribution. As these histograms differ so much, it means that their units of measure are not the same. In other words, for the same similarity/difference value, it may represents a 'huge' difference in one feature channel, but only a 'tiny' difference in the other channel. Therefore, it is necessary to standardize each feature channel before they are combined together.

Intuitively, the similarity distributions amongst the data points for a given dataset should not change with their representation features. As kernels measure the similarities between samples, we call this intuition the relative kernel distribution invariance (RKDI) property. Although there is no formal proof known to us at this stage, we believe it is a reasonable assumption and will show experimentally that maintaining such invariance can help improve performances.

We argue that before different kernels are linearly combined, the kernel values should be calibrated to a canonical feature space (CFS). In the absence of a known CFS, we use cross-validation to select one of the kernels as the CFS and calibrate all other kernels to this empirical CFS. This problem is the well-known histogram matching problem and our new feature combination framework is illustrated in Fig.1.

Let $HM(k_m(x,x'))$ represent the **H**istogram **M**atching operator that perform canonical histogram matching on the $m$-th kernel, then MKL and average kernel are represented as follows.

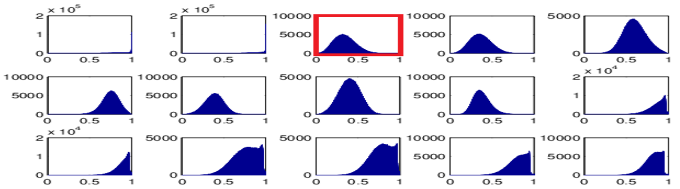

Figure 2: Kernel histograms of different features used in [3]. The histogram in the red box is chosen as the standard histogram.

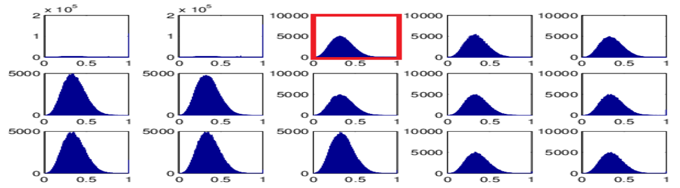

Figure 3: Histograms in Fig.2 after histogram matching.

The new average kernel $k^*$ is formed as:

$$k^*(x,x') = \frac{1}{F}\sum_{m=1}^{F} HM(k_m(x,x')) \tag{3}$$

In the case of MKL, the combined kernel $k^*$ is formed as:

$$k^*(x,x') = \sum_{m=1}^{F} \beta_m HM(k_m(x,x')) \tag{4}$$

We have performed extensive experiments on various computer vision and machine learning datasets and show that calibrating the kernels to an empirically chosen canonical space before they are combined can always achieve a performance gain over state-of-art methods. As histogram matching is a remarkably simple and robust technique, the new method is universally applicable to kernel-based feature combination.

The source code to reproduce the results reported in this paper will be made available on the second author's homepage.

[1] Francis R. Bach, Gert R. G. Lanckriet, and Michael I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. *ICML '04*, page 6, 2004.

[2] Peter Gehler and Sebastian Nowozin. On Feature Combination for Multiclass Object Classification. In *ICCV*, 2009.

[3] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. *ICCV*, September 2009.

[4] Francesco Orabona, Luo Jie, and Barbara Caputo. Online-Batch Strongly Convex Multi Kernel Learning. In *CVPR*, 2010.

[5] Manik Varma and Debajyoti Ray. Learning The Discriminative Power-Invariance Trade-Off. *ICCV*, October 2007.

[6] Andrea Vedaldi, Varun Gulshan, Manik Varma, and Andrew Zisserman. Multiple Kernels for Object Detection. *ICCV*, 2009.

[7] Fei Yan, Krystian Mikolajczyk, Mark Barnard, Hongping Cai, and Josef Kittler. lp Norm Multiple Kernel Fisher Discriminant Analysis for Object and Image Categorisation. In *CVPR*, 2010.