Class Consistent Multi-Modal Fusion with Binary Features

Ashish Shrivastava, Mohammad Rastegari, Sumit Shekhar, Rama Chellappa, Larry S. Davis
University of Maryland, College Park

Combining information from multiple sources - multiple sensor modalities or multiple feature channels applied to a single sensor modality - is generally advantageous for recognition problems. Fusing multiple modalities for classification has been explored in various computer vision applications [2, 3]. Many existing recognition algorithms combine different modalities based on training accuracy but do not consider the possibility of noise at test time. We present methods to perturb the test features so that all modalities agree on a common class label. We call this approach *class consistent multi-modal* (CCMM) fusion. We enforce that this perturbation be as small as possible via a quadratic program (QP) for continuous features, and a mixed integer program (MIP) for binary features. To efficiently solve the MIP, we utilize a greedy algorithm and empirically show that its solution is very close to that of a state-of-the-art MIP solver.

The key idea, summarized in Fig. 1, is to minimize the magnitude of perturbations to feature values for each modality to reach a point where all the modalities are predicting a common class label. This is formally posed as an optimization problem which minimizes the perturbation to satisfy the constraint that all modalities predict the same class label. Next, we establish our notation and develop the algorithm, first for continuous features and then, for binary features.

Assume that there are M modalities each with N_m labeled samples where $m = 1,\ldots,M$. Let the data matrix of the m^{th} modality be denoted by $\mathbf{Y}^{(m)} \in \mathbb{R}^{d\times N_m}$, where each column of $\mathbf{Y}^{(m)}$ is a d-dimensional data sample denoted by $\mathbf{y}_i^{(m)} \in \mathbb{R}^d$, for $i = 1,\ldots,N_m$. Let the class label of the i^{th} sample in the m^{th} modality be denoted by $l_i^{(m)} \in \{1,\ldots,C\}$, where C is the number of classes. Note that, for now, we regard the features as continuous; subsequently we adapt our method for binary features.

Let $\mathbf{W}^{(m)} := \begin{bmatrix} \mathbf{w}_1^{(m)} \\ \vdots \\ \mathbf{w}_C^{(m)} \end{bmatrix}$, be the classifier matrix for all categories in modality m, where the c^{th} row vector $\mathbf{w}_c^{(m)} \in \mathbb{R}^{1\times d}$ denotes the parameters of a linear classifier for the c^{th} class, which we refer to as a classification weight vector. These weight vectors are learned in a way that the class of a test sample $\mathbf{y}_p^{(m)}$ can be computed as,

$$\text{class of } \mathbf{y}_p^{(m)} = \arg\max_c \mathbf{w}_c^{(m)} \mathbf{y}_p^{(m)}. \tag{1}$$

In our implementation, we use an SVM [1] to learn these classification weight matrices $\mathbf{W}^{(m)}$ for all modalities.

Here, we describe the method for two modalities. Its extension to multiple modalitie is discussed in the paper. Denote a test sample's two modalities by $\mathbf{y}_p^{(1)}$ and $\mathbf{y}_p^{(2)}$ which by construction belong to the same class. Our goal is to minimize the total perturbation needed to reach the condition that the predicted classes using SVM matrices $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ are identical. This is captured in the following optimization problem,

$$\begin{aligned} &\min_{\mathbf{y}^{(1)},\mathbf{y}^{(2)}} \|\mathbf{y}^{(1)} - \mathbf{y}_p^{(1)}\|_2 + \|\mathbf{y}^{(2)} - \mathbf{y}_p^{(2)}\|_2 \\ \text{subject to,} \quad &\arg\max_c \mathbf{w}_c^{(1)} \mathbf{y}^{(1)} = \arg\max_c \mathbf{w}_c^{(2)} \mathbf{y}^{(2)} \end{aligned} \tag{2}$$

The optimization problem in (2) is non-smooth and non-convex due to the argmax functions. In order to solve it efficiently, we approximate it with a tractable convex problem. To achieve this, we employ an alternating optimization approach. First, we assume that the class predicted by the second modality is correct and optimize for $\mathbf{y}^{(1)}$, and then, we fix the class to the one predicted by the first modality and optimize for $\mathbf{y}^{(2)}$. When optimizing for the m^{th} modality feature $\mathbf{y}^{(m)}$, the class that is assumed to be correct is called the target class and is denoted by t_m. We seek to perturb the feature $\mathbf{y}_p^{(m)}$ so that its predicted class is t_m, which can be achieved by solving the following problem:

$$\begin{aligned} &\min_{\mathbf{y}^{(m)}} \|\mathbf{y}^{(m)} - \mathbf{y}_p^{(m)}\|_2 \\ \text{subject to,} \quad &\arg\max_c \mathbf{w}_c^{(m)} \mathbf{y}^{(m)} = t_m. \end{aligned} \tag{3}$$

As explained in the paper, the optimization problem in (3) is a quadratic program (QP) as following:

$$\begin{aligned} &\min_{\mathbf{y}^{(m)}} \|\mathbf{y}^{(m)} - \mathbf{y}_p^{(m)}\|_2 \\ \text{subject to,} \quad &\mathbf{A}_{t_m} \mathbf{y}^{(m)} \le 0, \end{aligned} \tag{4}$$

where matrix $\mathbf{A}_{t_m} \in \mathbb{R}^{C-1\times d}$ depends on the classification matrix $\mathbf{W}^{(m)}$ and the target class t_m. Let the solution of (4) be denoted by $\tilde{\mathbf{y}}_t^{(m)}$. Finally, the consistent class, denoted by l_p, across both modalities is the target class of the modality that requires the smallest change with respect to the original feature norm, i.e., $l_p = t_{m^*}$, where

$$m^* = \arg\min_m \frac{\|\tilde{\mathbf{y}}_p^{(m)} - \mathbf{y}_p^{(m)}\|_2}{\|\mathbf{y}_p^{(m)}\|_2}. \tag{5}$$

For binary features $\mathbf{b}_p^{(m)}$, the problem can be re-written with an additional constraint that the solution space should be binary:

$$\begin{aligned} &\min_{\mathbf{b}^{(m)}} \|\mathbf{b}^{(m)} - \mathbf{b}_p^{(m)}\|_1 \\ \text{subject to,} \quad &\mathbf{A}_{t_m} \mathbf{b}^{(m)} \le 0, \qquad \mathbf{b}^{(m)} \in \{0,1\}^d. \end{aligned} \tag{6}$$

As discussed in the paper, we formulate this problem into an MIP and provide an efficient greedy algorithm.

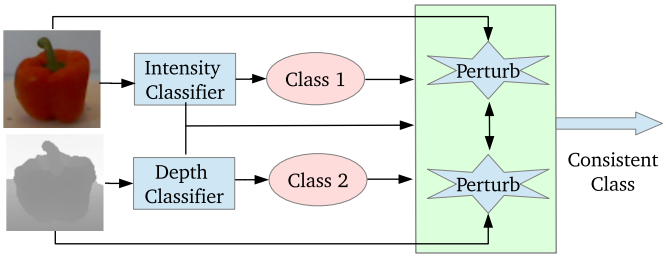

Figure 1: Overview of the proposed Class Consistent Multi-Modal (CCMM) fusion. The proposed algorithm perturbs the input features until all the modalities predict a consistent class.

[1] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[2] Dong Liu, Kuan-Ting Lai, Guangnan Ye, Ming-Syan Chen, and Shih-Fu Chang. Sample-specific late fusion for visual category recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[3] S. Shekhar, V.M. Patel, N.M. Nasrabadi, and R. Chellappa. Joint sparse representation for robust multimodal biometrics recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):113–126, Jan 2014.

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