A Multiple Kernel Learning Approach to Multi-Modal Pedestrian Classification

Marco San-Biagio[1], Aydın Ulaş[2], Marco Crocco[1],
Marco Cristani [1,2], Umberto Castellani[2], Vittorio Murino[1]
[1] *Pattern Analysis & Computer Vision, Istituto Italiano di Tecnologia, Via Morego, 30, 16163 Genova*
[2] *Department of Computer Science, University of Verona, Verona - Italy*
[1] *name.surname@iit.it,* [2]*name.surname@univr.it*

Abstract

Pedestrian detection is a key problem in many computer vision applications, especially in surveillance and security systems. To this end, information integration from different imaging modalities, such as thermal infrared and visible spectrum, can significantly improve the detection rate in respect to mono-modal strategies. For this reason, an effective fusion scheme is necessary to combine the information presented by multiple sensors. In this paper, we propose a pedestrian classification method based on the multiple kernel learning framework; standard pixel features (such as spatial derivatives) from both imaging modalities are employed to learn several feature-related basic kernels and a compound kernel is found as an optimized linear combination of basic kernels. Finally the compound kernel is used to train an SVM. Experiments performed on the OTCBVS dataset [1], demonstrate that our recipe definitely outclasses a wide set of literature fusion modalities.

1. Introduction

Detecting people is one of the most important task in Computer Vision [4, 5]. In particular, in the context of video surveillance, pedestrians are a very important, and very hard, class of objects to detect, especially in still images. This is because of the variability of pedestrian images: pedestrians can appear under different poses, different textures and different positions of the body articulation. For these reasons, many different approaches have been proposed in the literature: the standard strategy consists in the use of a sliding window approach over the whole image and the use of a classifier to evaluate each window. The classifier may be trained with different features, e.g., Haar-like features, Histograms of Oriented Gradients (HOG) edgelets or binary descriptors [4]. The most widespread classifiers are Support Vector Machines (SVMs, typically linear or RBF kernel SVMs)and boosting algorithms (e.g. AdaBoost, LogitBoost).

Over recent years, the researchers' attention focused on new imaging modalities to improve the reliability of the automated surveillance systems. In particular, thermal imaging is able to efficiently cope with working conditions that limit the use of visible imaging devices, such as night-time or adverse weather. Furthermore, thermal imaging is less affected by lighting conditions and provides enhanced contrast between human bodies and their environment. The above described approaches for pedestrian detection have been successfully transferred into the realm of thermal imaging [11].

Usually, the most pedestrian detection approaches rely on single-modality images. For this reason, finding an intelligent fusion of the information provided by both sensors reduces detection errors, thereby increasing the performance of tracking and the robustness of the surveillance system. Multi-modality fusion techniques applied to images, can be classified according to the level of processing where the fusion takes place. In particular, three are the levels [7]: *Pixel Level*, *Feature Level* and *Decision Level*. *Pixel level* fusion means the fusion at the lowest processing level, referring to the merging of the physical parameters of the source images. Fusion at the *feature level* implies the concatenation of two feature sets, separately extracted from the two image sources, into a unique feature vector which is then fed into a classifier. Finally, *decision level* methods are at the highest level of processing where information fusion can take place. In this class, the decisions of each single classifier are fused in order to generate the final decision, according to a given policy.

This paper presents a novel approach aimed at fusing the information provided by thermal infrared and visible spectrum images for robust pedestrian classification through the use of Multiple Kernel Learning framework. Selecting the kernel function and its parameters is an important issue in training a classifier. Besides standard kernels such linear or RBF ones, several kernel functions specifically targeted to particular applications, such as natural language processing [9] and bioinformatics [10], have been proposed in literature. In MKL a compound kernel is learned from a combination of different basic kernels on the base of following considerations: (a) different kernels correspond to different notions of similarity and instead of trying to find which works best, a learning method does the picking for us, or may use a combination of them; (b) different kernels may use inputs coming from different sources or modalities. For a general survey on MKL see [6].

In this paper we propose to apply MKL as a new fusion policy aimed at improving classification performance yielded by single imaging modalities. In particular each basic kernel is learned from a dense feature map taken from either visible or thermal image. Subsequently, according to the two stage approach proposed in [3] a compound kernel is built as a weighted sum of basic kernels, where the weights are optimized in order to maximize the alignment between the compound kernel and the ideal kernel given by the product of the labels. Finally a single SVM is trained using such compound kernel. The fusion performed at the *kernel level* can be seen as an intermediate level between *feature-level* and *decision level*, allowing to overcome the drawbacks related to both modalities. In particular, as each feature map is related to just one basic kernel, *kernel-level* fusion avoids the sparsity of data which often occurs in the combined feature space given by *feature-level* fusion. On the other hand, all the feature set is involved in kernel optimization, so limiting the loss of information in the transition from data to single decisions typical of *decision-level* fusion policies. Moreover, the optimized weights of the basic kernels may provide information on the relative importance of each feature and each sensory modality for the classification task.

The proposed method has been tested on OCTBVS database [1] and compared with both feature level and decision level standard approaches, showing a definite improvement over all the methods tested. The rest of the paper is organized as follows: in Section 2, the proposed method is described, in section 3, results are discussed and compared, finally in section 4 some conclusions are drawn.

2. Proposed method

In a standard SVM classifier the decision function is given by [12]:

$$d(\vec{x}) = \sum_{i=1}^{N} \alpha_i y_i k(\vec{x}_i, \vec{x}) + b \tag{1}$$

where $\vec{x}$ is the feature vector, $y_i \in \{-1, +1\}$ are the training labels, α_i are the coefficient learned in the training phase and $k(\vec{x}_i, \vec{x})$ is the kernel function (enconding similarity between data instances). In this case, the kernel function is *a priori* fixed. On the contrary, MKL methods [2, 8] learn a combination k_η of multiple kernels:

$$k_\eta(\vec{x}_i, \vec{x}_j; \vec{\eta}) = f_\eta(\{k_m(\vec{x}_i^m, \vec{x}_j^m)_{m=1}^M\}; \vec{\eta}) \tag{2}$$

where the combination function f_η forms a single kernel from M basic kernels k_m, each one related to a generally different feature vector $\vec{x}^m$, using the weights $\vec{\eta}$. This allows one to perform the selection/combination of different kernels or data sources automatically. There is significant work on the theory and application of MKL, and for most algorithms, the difference is the optimization method which is applied to estimate the weights or the combination rule used [2, 8, 6, 3]. In this work we apply one of the *linear*-MKL methods [2, 8], where the general formulation with M kernels can be defined as:

$$k_\eta(\vec{x}_i, \vec{x}_j; \vec{\eta}) = \sum_{m=1}^{M} \eta_m k_m(\vec{x}_i^m, \vec{x}_j^m) \tag{3}$$

with $\eta_m \in \mathbb{R}$. To optimize weights $\vec{\eta}$ the approach proposed in [3] is adopted. Such an approach is based on the maximization of a measure of similarity between kernel matrices called *centered-kernel alignment* (CA) defined as follows:

$$\mathrm{CA}(\mathbf{K}_1, \mathbf{K}_2) = \frac{\langle \mathbf{K}_1^c, \mathbf{K}_2^c \rangle_F}{\sqrt{\langle \mathbf{K}_1^c, \mathbf{K}_1^c \rangle_F \langle \mathbf{K}_2^c, \mathbf{K}_2^c \rangle_F}}$$

where $\mathbf{K}$ is the generic kernel matrix, $\mathbf{K}^c$ is the centered version of $\mathbf{K}$ [3] and F is the Frobenius norm. In particular, the following constrained maximization problem is solved [3]:

$$\begin{aligned} &\text{maximize } \mathrm{CA}(\mathbf{K}_\eta, \vec{y}\vec{y}^\top) \\ &\text{with respect to } \vec{\eta} \in \mathcal{M} \end{aligned} \tag{4}$$

where $\mathcal{M} = \{\vec{\eta}\colon \|\vec{\eta}\|_2 = 1 \wedge \vec{\eta} \in \mathbb{R}_+^M\}$ and $\vec{y}\vec{y}^\top$ is *the ideal kernel* for the classification task. It is demonstrated [3] that if $\vec{v}^*$ is the solution of the following Quadratic Programming problem

$$\begin{aligned} &\text{minimize } \vec{v}^\top \mathbf{M} \vec{v} - 2\vec{v}^\top \vec{a} \\ &\text{with respect to } \vec{v} \in \mathbb{R}_+^M \end{aligned} \tag{5}$$

the solution to (4) is given by $\vec{\eta}^* = \vec{v}^*/\|\vec{v}^*\|_2$, where $\mathbf{M} = \{\langle \mathbf{K}_m^c, \mathbf{K}_h^c \rangle_F\}_{m,h=1}^M$ and $\vec{a} = \{\langle \mathbf{K}_m^c, \vec{y}\vec{y}^\top \rangle_F\}_{m=1}^M$. Finally, the compound kernel resulting from the optimized weights $\vec{\eta}^*$ is employed to train a single SVM.

In our setup, each feature vector $\vec{x}^m$ is built from a dense map of a single low level feature extracted either from a visible spectrum image or its thermal counterpart. In particular, let us assume that *V* is a visible spectrum gray scale image and *T* a thermal gray scale image of the same size of P pixels, portraying the same scene. *V* and *T* are registered in order to have a pixel-to-pixel correspondence, and they contain either a pedestrian or a portion of background. For each pixel p of each visible image, a set of low level features is extracted:

$$\vec{v}_p = \begin{bmatrix} V_x & V_y & V_{xx} & V_{yy} & \sqrt{V_x^2 + V_y^2} & LBP(V) \end{bmatrix} \tag{6}$$

where V_x, V_{xx}, etc. are first- and second-order derivatives of the visible image intensity, and the last term represents the Local Binary Pattern [13], a descriptor encoding local texture information around the given pixel. Similarly, for the corresponding thermal image *T* and on the same pixels a vector of low level features $\vec{t}_p$ is extracted as:

$$\vec{t}_p = \begin{bmatrix} T_x & T_y & T_{xx} & T_{yy} & \sqrt{T_x^2 + T_y^2} & LBP(T) \end{bmatrix} \tag{7}$$

The two vectors are arranged in a single visible-thermal vector as:

$$\vec{f}_p = \begin{pmatrix} \vec{v}_p & \vec{t}_p \end{pmatrix} \tag{8}$$

Finally, the feature vectors $\vec{x}^m$ related to the M basic kernels are built as:

$$\vec{x}^m = \begin{bmatrix} f_{1m} & f_{2m} & \cdots & f_{Pm} \end{bmatrix} \tag{9}$$

where f_{pm} is the m-th value of vector $\vec{f}_p$.

2.1 Compared methods

The proposed fusion method is compared with the standard *feature-level* method and five different *decision-level* methods. In particular, denoting with d_x the final decision function and with d_x^m the decision function given by a single kernel SVM trained on the m-th feature vector $\vec{x}^m$, the following methods have been compared:

- `MKL`: d_x is given by the proposed method
- `SVM`: $d_x = d_x^{m^*}$, where i^* is the single best linear kernel without any combination
- `AVG`: $d_x = (\sum_{m=1}^N d_x^m)/N$. This is the average rule
- `MED`: $d_x = \mathrm{med}(d_x^m), m = 1 \ldots N$, where med is the median operator. This is the median rule of classifier combination
- `VOT`: $d_x = \sum \mathrm{sgn}(d_x^m), m = 1 \ldots N$, where sgn is the signum operator. This is the majority voting
- `MAX`: $d_x = \max|(d_x^m)| * \mathrm{sgn}(\max(d_x^m)), m = 1 \ldots N$. This is the max rule of classifier combination
- `CON`: d_x is obtained by concatenation all N feature sets and training a single linear SVM

3. Results

In this section, we report experimental results on data extracted from the OTCBVS Benchmark Dataset [1]. In particular, we picked up two datasets from Ohio State University Color-Thermal Database, composed of a sequence of frames, at 8-bit gray scale format, taken from two couples of fixed cameras (thermal and visible) in two different urban outdoor locations. From each sequence we manually extracted a number of windows, enclosing either pedestrians or portions of background, which were resized at 36x18 pixels. The two final datasets were composed of 7072 windows (3536 pedestrians and 3536 non-pedestrians) for the first dataset, and 5236 windows (2618 pedestrians and 2618 non-pedestrians) for the second dataset. In Figure 1 some examples of windows from the two datasets are shown. The two datasets will be made publicly available in order to encourage further research on this topic.

Figure 1. Pedestrian and Background examples extracted from the two datasets.

From each image we extracted twelve different feature maps of size equal to the original images. We decided to perform training on the first dataset and testing on the second one and *viceversa*. In such a way, training and testing background examples are sufficiently different so as to guarantee that the obtained performance is representative of a general situation where background is not *a priori* known . At the same time, the two background scenes belongs to the same typology (urban outdoor environment), and are sufficiently similar so as to

avoid skewing results. In Figure 2 the classification performance of the different methods are shown in terms of Detection Trade-Off (DET) curves, defined as the Miss Rate $\left(\frac{\#FalseBackground}{\#TotalPedestrians}\right)$ versus False Positives Rate (FPRate) $\left(\frac{\#FalsePedestrians}{\#TotalBackgrounds}\right)$. As can be seen, the proposed method outperform all the other fusion policies. As an example for a FPRate of 10^{-2} MKL provides a Miss Rate of $5 \cdot 10^{-3}$ against about $1.5 \cdot 10^{-2}$ for MED and about $2 \cdot 10^{-2}$ for VOT and CON. Furthermore, in figure 3 the weight distribution for each of the twelve features is drawn. In the figure each *v* identifies the visible features, instead each *t* identifies the thermal features. Interestingly, non-zero weights are picked up from both thermal and visible features, so confirming that complementary information for pedestrian detection is brought up by the two imaging modalities.

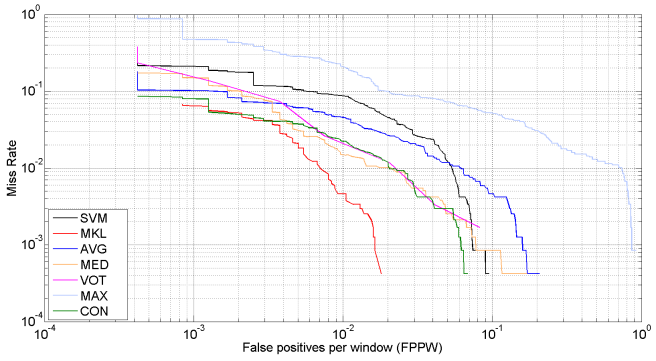

Figure 2. DET curve.

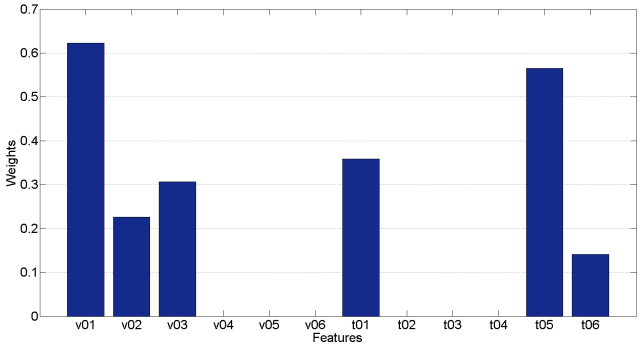

Figure 3. Weights distribution.

4. Conclusion and Future Works

This paper proposes a novel fusion policy for thermal and visible spectrum images, in a pedestrian classification context. The proposed method, based on a MKL framework, definitely outclasses a wide set of literature fusion policies on a visible-thermal image dataset. Future works will concentrate on the creation of an ad-hoc large size database with variable backgrounds (e.g. taken from a car moving in urban environment) in order to further validate the proposed method on different scenarios.

Acknowledgments. We want to thank Dr. Mehmet Gönen for the discussions on MKL.

References

[1] OTCBVS benchmark dataset collection. http://www.cse.ohio-state.edu/OTCBVS-BENCH/bench.html.

[2] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the twenty-first international conference on Machine learning, ICML '04*, pages 41–48, 2004.

[3] C. Cortes, M. Mohri, and A. Rostamizadeh. Two-stage learning kernel algorithms. In *ICML*, pages 239–246, 2010.

[4] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PP(99):1, 2011.

[5] M. Enzweiler and D. Gavrila. A multilevel mixture-of-experts framework for pedestrian classification. *Image Processing, IEEE Transactions on*, 20(10):2967 –2979, oct. 2011.

[6] M. Gönen and E. Alpaydïn. Multiple kernel learning algorithms. *J. Mach. Learn. Res.*, 12:2211–2268, July 2011.

[7] D. Hall and J. Llinas. An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1):6 –23, jan 1997.

[8] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, December 2004.

[9] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *J. Mach. Learn. Res.*, 2:419–444, Mar. 2002.

[10] B. Schlkopf, K. Tsuda, and J.-P. Vert. *Kernel methods in computational biology*. MIT Press, Cambridge, Mass., 2004.

[11] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi. Pedestrian detection using infrared images and histograms of oriented gradients. In *Intelligent Vehicles Symposium, 2006 IEEE*, pages 206 –212, 0-0 2006.

[12] V. N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, 1998.

[13] Y. Zheng, C. Shen, R. I. Hartley, and X. Huang. Effective pedestrian detection using center-symmetric local binary/trinary patterns. *CoRR*, abs/1009.0892, 2010.