ASAI 2015, 16º Simposio Argentino de Inteligencia Artificial.

Detección de Peatones Utilizando Optimización Multi-Objetivo

Pablo Negri[1,2]

[1] CONICET, Av. Rivadavia 1917, Buenos Aires, Argentina.
[2] INTEC-UADE, Lima 717, Buenos Aires, Argentina.

Abstract. La detección de peatones en secuencias de video urbanas representa un desafío para los sistemas de clasificación. Cuando el fondo de la escena donde circulan las personas es muy texturado, el desempeño de la mayoría de los clasificadores se ve severamente afectado. En este artículo se propone la utilización de una técnica de Optimización Multi-Objetivo (en inglés *Multi-Objective Optimization* o *MOO*). La metodología entrena un pool de Cascadas de Clasificadores Dopados a partir de diferentes conjuntos de aprendizaje, otorgándoles un comportamiento particular. El análisis de sus curvas ROC permite construir un frente de Pareto que selecciona los puntos operacionales localmente dominantes. Los resultados de esta metodología sobre una secuencia real muestran una mejora en la performance del sistema de detección.

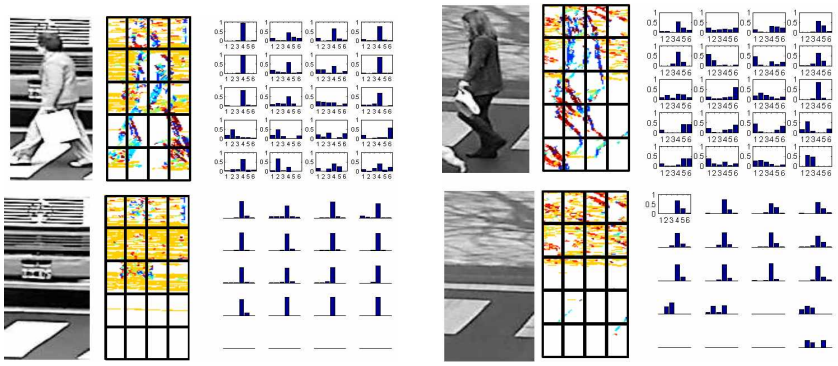
Keywords: Optimización Multi-Objetivo, Detección Peatones

1 Introducción

En la etapa de concepción de sistemas de procesamiento de imágenes aplicados a la detección de objetos es preciso fijar criterios para determinar su comportamiento. En la mayoría de los casos tendremos dos clases de respuestas: Detecciones Correctas (DC) y Falsas Alarmas (FA). Las DC indican los casos en los que la clase objeto fue exitosamente identificada en la imagen. Las FA son las respuestas erróneas del sistema.

Dependiendo la aplicación, el desempeño en DC del sistema puede ser fundamental para identificar todos los objetos/situaciones: un detector de minas personales enterradas en un campo debe ser muy sensible y responder positivamente ante cualquier duda. Esto implica que se obtengan valores muy altos de DC, al mismo tiempo que muchas FA. Dado que el precio a pagar por una mina personal no detectada puede ser una vida humana, no es muy importante el hecho de obtener una tasa elevada de FA. Otro ejemplo puede ser un sistema de fumigación rural que detecta maleza usando la visión. Un número elevado de FA redunda en una pérdida económica al pulverizar innecesariamente el sembrado. Un valor de FA cercano a cero significa que el sistema responde positivamente si está 100 % seguro, con lo cual no atacaría toda la maleza presente. Esto, sin embargo, puede no ser problemático: una cantidad pequeña de maleza no

(a) Secuencia Urbana



(b) Ejemplo 1 (c) Ejemplo 2

Fig. 1. La fig. (a) es una captura de la secuencia urbana en estudio. Las figs. (b) y (c) son ejemplos de peatones en fondos texturados no detectados por clasificadores.

representa peligro para la siembra. Determinar el comportamiento del sistema implica entonces la búsqueda de un equilibrio en el número de DC y FA ligado a la aplicación.

En este artículo se estudian diferentes relaciones entre DC y FA aplicados a la detección de personas. Las imágenes tomadas en la vía pública, como muestra la fig. 1(a), implican un escenario no controlado donde diversos factores atentan contra el buen desempeño de detectores del estado del arte [5, 7, 10]. Esto es debido principalmente a: otros objetos en movimiento en la escena (vehículos), cambios abruptos en la iluminación y un fondo muy texturado.

Las figs. 1(b) y 1(c) muestran dos ejemplos de peatones no detectados por los clasificadores. Este trabajo utiliza el Movement Feature Space (MFS) [9, 10] para describir la dinámica de la escena. El MFS es un espacio de descriptores que codifica en líneas de nivel orientadas los objetos que no pertenecen al fondo estático. La información de los objetos en movimiento es volcada en un plano de orientaciones O_t, mostrado en la fig. 1(b), donde cada color corresponde a un valor de orientación diferente. Los Histogramas de Líneas de Nivel (HO2L) son calculados en cada rectángulo de la grilla sobre O_t acumulando los pixeles con la misma orientación (por detalles remitirse a [10, 9]).

Los HO2L del ejemplo 1 del fondo sin la persona, fig. 1(b), contienen fundamentalmente líneas de nivel con orientación horizontal: $bin = 4$. Cuando la

persona está frente al vehículo, su presencia hace variar los HO2L anteriores, aunque existe una muy alta predominancia de la orientación horizontal. Podemos decir que sus HO2L están *sumergidos* dentro de los descriptores del fondo. El ejemplo 2 muestra un cambio de iluminación dado por un haz de luz que logra colarse entre las nubes, generando movimiento en el MFS y una configuración similar de los HO2L al ejemplo 1, aunque con un efecto menos marcado. Los peatones caminando delante de este fondo tampoco son detectados.

El problema abordado se relaciona con la presencia de descriptores horizontales que no son generalmente generados por una persona [7]. Esto convierte a esta orientación en un factor muy discriminante para los clasificadores a la hora de eliminar FA. Sin embargo, una persona sumergida en un fondo con estas características es ignorada como vimos anteriormente. Para limitar este efecto, se podría entrenar un clasificador agregando un gran número de imagenes de personas en fondos texturados, pero esto redunda en un aumento exponencial de las FA: las lineas horizontales ya no son tan discriminantes. Llegamos así a un problema con objetivos contradictorios para el cual es necesaria una solución diferente.

El artículo propone una técnica de Optimización Multi-Objetivo (MOO) que mejore la performance del sistema. La metodología consiste en entrenar un conjunto de clasificadores utilizando diversos grupos de entrenamiento cada vez más heterogéneos. A partir de sus curvas Receiver Operating Characacteristics (ROC) se construye un frente de Pareto usando los puntos operacionales localmente óptimos [4]. Trabajos recientes de la literatura [4, 8, 11] aplican MOO para seleccionar los mejores hiperparámetros de los clasificadores. En este artículo se busca elegir combinaciones de soluciones de Pareto que sean aplicadas dependiendo la dinámica de la escena.

La sección siguiente detalla la metodología de entrenamiento y obtención del pool de clasificadores. Los resultados del sistema se presentan y discuten en la sec. 3, para terminar luego en la sec. 4 con las conclusiones de este trabajo.

2 Metodología

Construcción del Frente ROC La optimización de Pareto permite dar un marco a la existencia de soluciones con objetivos contrapuestos. En este marco definimos Ψ como el posible conjunto de soluciones posibles que pueden obtenerse del entrenamiento de nuestro sistema. El $\mathbf{x} = [x_1, ..., x_n]^T \in \Re^n$ es un vector solución computesto de variables de decisión x_i, y $f_i(\mathbf{x}), i = 1, ..., l$ son las posibles l funciones objetivo. Luego, decimos que la solución $\mathbf{x}^1$ domina a la solución $\mathbf{x}^2$ ($\mathbf{x}^1 \leq \mathbf{x}^2$), si y solo si $\mathbf{x}^1$ es mejor que $\mathbf{x}^2$ en un objetivo y no es peor en el resto [11]: $\forall i : f_i(\mathbf{x}^1) \leq f_i(\mathbf{x}^2) \wedge \exists j : f_j(\mathbf{x}^1) < f_j(\mathbf{x}^2)$.

Utilizando este concepto de dominancia, el MOO encuentra un conjunto de todas las soluciones dominantes teniendo en cuenta las funciones objetivo de nuestro sistema. Este conjunto se denomina Frente de Pareto. En este trabajo, las funciones objetivos van a estar vinculadas a la relación FA/DC obtenidas desde la curva ROC [4]. Para construir la ROC de un clasificador bi-clases precisamos

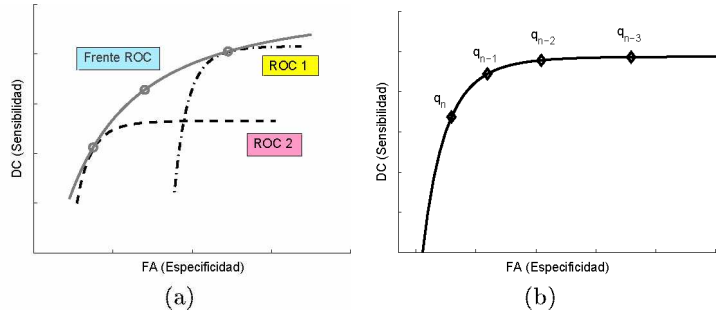


Fig. 2. La figura muestra: (a) el frente de Pareto obtenido utilizando dos curvas ROC, (b) la ROC de una Cascada de Clasificadores Dopados con los puntos operacionales q_i.

un conjunto de ejemplos positivos y negativos, obteniendo para cada uno un *score*. Aplicando un umbral de validación sobre estos *scores* se obtiene un valor de DC y de FA, *scores* positivos y negativos (respectivamente) por encima del umbral. Estos dos valores representan un punto en la ROC. Haciendo variar el umbral de validación se obtiene finalmante la curva [3]. Esta curva nos permite evaluar la sensibilidad y especificidad del clasificador.

La fig. 2(a) representa a modo de ejemplo dos curvas ROC pertenecientes a dos clasificadores diferentes. Definimos las funciones objetivo f maximizando las DC y minimizando las FA. La ROC1 domina localmente a la ROC2 para valores superiores de DC y viceversa. La selección de puntos operacionales dominantes localmente define el Frente ROC de Pareto, que se representa en la fig. 2(b) como la curva gris exterior, más cercana a la esquina superior izquierda del plano ROC (DC=1, FA=0).

Las subsecciones siguientes combinan el entrenamiento de las Cascadas de Clasificadores Dopados, con la técnica de Optimización Multi-Objetivo.

Cascada de Clasificadores Dopados El entrenamiento de una Cascada de Clasificadores Dopados $\mathbf{C}$ utilizando el algorithm de Adaboost precisa una base de ejemplos positivos P, una base de ejemplos negativos N. También es necesario definir parámetros de aprendizaje: el número de etapas máximo E de la cascada, y los porcentajes de detecciones mínimas d_{min} y falsas alarmas máximas f_{max} por etapa [12].

La Cascada resultante $\mathbf{C} = \{C_1, C_2, ..., C_n\}$ es una serie de n clasificadores C_i de complejidad creciente. Los C_i son aplicados sucesivamente a regiones x_k de una imagen de test para detectar el objeto de interés (peatones en nuestro caso): $C_i(x_k) = s_k$. Si $s_k > T_i$, donde T_i es el umbral de C_i obtenido en el entrenamiento, x_k será tratada por el siguiente clasificador C_{i+1}.

La ROC de $\mathbf{C}$ se calcula siguiendo la técnica propuesta por Viola [12]. La curva total se obtiene concatenando segmentos de ROC calculados para cada clasificador de $\mathbf{C}$. Un segmento j de la curva corresponde al clasificador C_j es resultado de aplicar un umbral de validación sobre los scores obtenidos en el

conjunto de test, desde $-\infty$ hasta el valor de T_j. La curva 2(b) muestra un ejemplo con q_n, indicando el punto operacional del último clasificador C_n de $\mathbf{C}$, q_{n-1} el punto operacional del clasificador C_{n-1}, etc.

Entrenamiento Multi-Objetivo El comportamiento de $\mathbf{C}$ está relacionado con la elección de las variables del entrenamiento $\{P, N, E, d_{min}, f_{max}\}$. En este trabajo se estudia la sensibilidad a la naturaleza de los ejemplos positivos que componen la base de entrenamiento P. Si P es muy homogénea, durante el entrenamiento de $\mathbf{C}$ los ejemplos positivos son proyectados en un espacio de clasificación donde son fácilmente agrupados en núcleos. Ejemplos disimiles a la clase promedio son proyectados lejos de estos núcleos y pueden ser considerados como negativos. El desempeño de este tipo de clasificadores resulta en muy bajas FA, aunque un porcentaje de DC no muy elevado. Una curva ROC correspondiente a este tipo de Cascadas es la ROC2 de la fig. 2(a). Para un P heterogéneo, Adaboost requiere más procesamiento para agrupar los ejemplos en el espacio de clasificación, generando fronteras mucho más abiertas. Este tipo de clasificadores posee número elevado de DC, a costa de un incremento de las FA. La curva ROC1 de la fig. 2(a) ejemplifica su comportamiento frente a ROC2.

Se prone el algoritmo 1 que entrena un conjunto $\mathcal{N}$ de Cascadas de Clasificadores Dopados: $\mathcal{C}_{MOO} = \{\mathbf{C}_1, ..., \mathbf{C}_{\mathcal{N}}\}$. La función $EntrenarCascada()$ de la línea 4 aplica la metodología de [12] para entrenar todos los C_i, usando como argumento de entrada el conjunto de ejemplos positivos P_k, cuya heterogeneidad crece en cada iteración de i de 1 a $\mathcal{N}$. De esta forma, $C_{\mathcal{N}}$ poseerá frenteras mucho más *amplias* en el espacio de clasificación que C_1. El aumento de heterogeneidad se logra eliminando de P_k los ejemplos positivos ubicados en los centros de los nucleos en el espacio de clasificación generado por C_k. Las funciones $EvaluarScoresSalida()$ y $OrdenarIndicesEnFormaCreciente()$ obtienen y ordenan los scores de todos los ejemplos en P_k usando C_k. Los más altos scores se ubicarán al final de la lista de índices $oidx$. La variable p, que representa el número total de positivos en P_k se reduce en la línea 8 por un factor d_{min}^n (con $d_{min} < 1$). Es creado un nuevo conjunto de positivos P_{new} tomando los primeros p elementos correspondientes a la lista ordenada $oidx$. P_{new} será el nuevo conjunto de positivos que entrene C_{k+1}.

3 Experimentos y Resultados

Bases de Entrenamiento y Validación La base positiva $\mathcal{P}$ consiste en imágenes rectangulares conteniendo una persona extraídas de capturas de video como las mostradas en la fig. 1(a). El número todal de ejemplos es duplicado al invertir los parches a partir de su eje vertical obteniendo 6.726 positivos.

La base de imágenes negativas (sin personas) empleada para el entranamiento es el PASCALVOC 2012 (7.166 imágenes) [2]. Para construir las curvas ROC y el frente de Pareto se utiliza el INRIA person negative (1.570 imágenes) [1].

El sistema de detección de peatones es aplicado en los GSDatasets, que son dos secuencias urbanas de dos minutos de duración capturando la vista frontal

Algorithm 1 Entrenamiento Multi-Objetivo

Require: P base positiva, N base de negativos, y $\{E, d_{min}, f_{max}\}$
Ensure: Set de Cascadas Multi-Objetivo $\mathcal{C}_{MOO}$
1: $k \leftarrow 1$
2: $P_k \leftarrow P$, $p \leftarrow$ # positivos en P
3: **while** $k < \mathcal{N}$ **do**
4: $\mathbf{C}_k \leftarrow$ EntrenarCascada($P_k, N, E, d_{min}, f_{max}$)
5: $n \leftarrow$ # de etapas en ($\mathbf{C}_k$)
6: $s, idx \leftarrow$ EvaluarScoresSalida($\mathbf{C}_k, P_k$)
7: $oidx \leftarrow$ OrdenarIndicesEnFormaCreciente(s, idx)
8: Actualizar p: $p \leftarrow ((d_{min})^n \cdot p)$
9: Crear P_{new} como los primeros p elementos de P_k ordenados por $oidx$
10: Guardar $\mathbf{C}_k$ en $\mathcal{C}_{MOO}$
11: $k \leftarrow k + 1$
12: $P_k = P_{new}$
13: **return** $\mathcal{C}_{MOO}$

de una esquina donde peatones cruzan la calle. Los datasets son públicos y están disponibles en `http://pablonegri.free.fr/Downloads/GSdataset-PANKit.htm`. En ambas secuencia, para cada uno de los peatones visibles se definió un rectángulo englobante y una etiqueta única permitiendo evaluar métodos de seguimiento. En la secuencia GS06, 5 personas cruzan la calle y generan 1.157 posiciones a detectar por el clasificador. La secuencia GS54 tiene 3.644 posiciones a detectar generadas por 16 personas que cruzan la calle.

Entrenamiento de los sets $\mathcal{C}_{MOO}$ En el aprendizaje se utiliza la técnica 3 fold-cross validation. La base positiva $\mathcal{P}$ completa se divide al azar en tres $\{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3\}$. Cada entrenamiento de $\mathcal{C}_{MOO}$ utiliza $P = \mathcal{P}_i \cup \mathcal{P}_j$, y el $\mathcal{P}_k$ restante, denominamo V, sirve para calcular la curva ROC que caracterice el comportamiento del $\mathcal{C}_{MOO}$.

Para cada $\mathcal{C}_{MOO}$ se entrenaron $\mathcal{N} = 5$ Cascadas $\mathbf{C}_i$ siguiendo el algoritmo 1. Esta cantidad fue elegida en base al número de ejemplos positivos que van descartándose en cada iteracion, que para el entrenamiento de $\mathbf{C}_5$ puede descender hasta uno 3.253 ejemplos positivos. Las variables d_{min} y f_{max}, en nuestros experimentos, toman los valores $0,995$ y $0,4$ respectivamente.

La fig. 3 muestra las curvas ROCs obtenidas de un set $\mathcal{C}_{MOO}$, utilizando el dataset V (2.242 ejemplos positivos) y 100.000 parches negativos del INRIA negative dataset tomados al azar. En la fig. 3 fueron graficadas las ROC de $\mathbf{C}_1$ y $\mathbf{C}_5$. El frente de Pareto puede ser estimado eligiendo los puntos operacionales de una de las ROC que dominen localmente los puntos operacionales de la otra ROC, maximizando los DC y minimizando las FA.

Implementación del sistema $\mathcal{C}_{MOO}$ En esta sección se propone una metodología para seleccionar automáticamente la Solución Óptima de Pareto que dependa de la dinámica de la escena.

De la fig. 3, el Frente de Pareto se compone de los puntos operacionales dominantes localmente pertenecientes al clasificador C_1 o C_5. Ante un cambio de la dinámica de la escena, es posible elegir un punto operacional del Frente ROC que pertenezca a un clasificador o al otro.

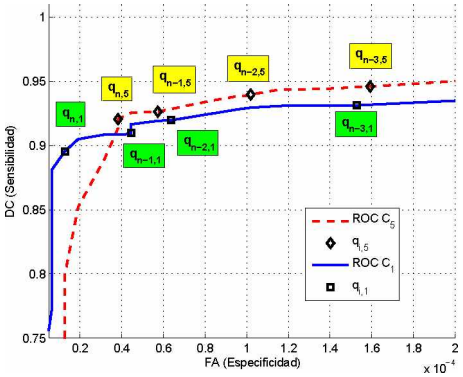


Fig. 3. Se muestran dos curvas ROC obtenidas para $\mathbf{C}_1$ y $\mathbf{C}_5$ y los puntos operacionales $q_{i,1}$ de la cascada $\mathbf{C}_1$ y los $q_{i,5}$ correspondientes a la cascada $\mathbf{C}_5$.

Para simplificar la operación, dos puntos operacionales del Frente ROC de Pareto son aplicados dependiendo del estado del semáforo. Debido a que se persigue el incremento de las DC cuando las personas se desplazan delante de los vehículos detenidos, el clasificador elegido debe tener fronteras más amplias en el espacio de clasificación. Sin embargo, el uso continuo de este punto operacional, que tiene también un número más importante de FA, empeora el desempeño del sistema. Cuando las personas no cruzan más debido a que los autos circulan por la calle, la Solución de Pareto podría cambiar a un punto operacional perteneciente a un clasificador con fronteras más estrechas en el espacio de clasificación. Dado que la dinámica de los vehículos está gobernada por el semáforo, el cambio de punto operacional podría estar determinado por sus estados:
- Semáforo Rojo: punto operacional $q_{rojo} \rightarrow \mathbf{C}_5\{q_{n,5}\}$ o $\mathbf{C}_5\{q_{n-1,5}\}$, o $\mathbf{C}_5\{q_{n-2,5}\}$, para maximizar las DC.
- Semáforo Verde: punto operacional $q_{verde} \rightarrow \mathbf{C}_1\{q_{n,1}\}$, para minimizar las FA.

El estado del semáforo de los vehículos se infiere del estado de los peatones visible en la escena (ver fig. 1).

En la tabla siguiente presentamos los resultados aplicando el sistema MOO. Los resultados son comparados con una Cascada de Clasificadores clásica $\mathbf{C}$, que corresponde al caso de no utilizar el sistema multi-objetivo. El desempeño se evalúa usando: error del sistema ($missrate$) que marca el porcentaje de peatones que no fueron detectados, el *false positive per image* (FPPI) calculado como el promedio de falsas alarmas por imagen en cada el dataset, y el *average precision ratio* (AP) obtenido de la curva *Precision-Recall* para el punto operacional elegido. Estos indicadores son habitualmente empleados para comparar los sistemas de detección [6].

Los resultados muestran que el $\mathcal{C}_{MOO}\{\{q_{n,1}, q_{n,5}\}$ minimiza el *miss rate* un 10%, manteniendo un número similar de FA. Esto hace que el AP sea máximo para este caso, lo cual lo puede convertir en el sistema ideal para esta aplicación.

	GS06			GS54		
Detector	Miss Rate (%)	FPPI	AP	Miss Rate (%)	FPPI	AP
C	28.92	0.55	68.4	35.03	1.02	60.39
$\mathcal{C}_{MOO}\{q_{n,1}, q_{n,5}\}$	**25.98**	0.57	**69.61**	**32.85**	1.04	62.46
$\mathcal{C}_{MOO}\{q_{n,1}, q_{n-1,5}\}$	26.86	0.81	69.21	31.72	1.25	**62.63**
$\mathcal{C}_{MOO}\{q_{n.1}, q_{n-2.5}\}$	26.82	0.99	68.84	31.09	1.51	62.15

Table 1. Resultados de la detección de los diversos sistemas.

Como era de esperar, al tomar puntos operacionales $q_{n-1,5}$ y $q_{n-2,5}$ mejoran las detecciones pero también aumentan el FPPI.

4 Conclusiones

El artículo propone un sistema de Optimización Multi-Objetivo aplicado a la detección de personas en escenas urbanas de alta complejidad. El conjunto $\mathcal{C}_{MOO}$ de clasificadores C_i se entrena eligiendo diferentes conjuntos de ejemplos positivos. Dependiendo en la dinámica de la escena, diferentes puntos operacionales dominantes localmente y que componen un Frente de Pareto son aplicados para mejorar el desempeño del sistema. Las perspetivas y trabajos futuros estarán orientados a desarrollar una metodológica que optimice la elección de ejemplos positivos para la base de entrenamiento.

Agradecimientos Financiaron este trabajo la ACyT A14T24 (UADE) y el PICT-BICENTENARIO 2283 (FONCYT).

References

[1] INRIA Person dataset (junio 2015), `http://pascal.inrialpes.fr/data/human/`
[2] PASCAL VOC2012 dataset (junio 2015), `http://host.robots.ox.ac.uk:8080/`
[3] Bradley, A.: The use of the area under the roc curve in the evaluation of machine learning algorithms. PR 30, 1145–1159 (1997)
[4] Chatelain, C., et al.: A multi-model selection framework for unknown and/or evolutive misclassification cost problems. PR 43(3), 815–823 (2010)
[5] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. vol. 1, pp. 886–893 (2005)
[6] Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. PAMI 34(4), 743–761 (2012)
[7] Felzenszwalb, P., Girshick, G., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI 32(9), 1627–1645 (2010)
[8] Li, W., Liu, L., Gong, W.: Multi-objective uniform design as a svm model selection tool for face recognition. Expert Systems with Applications 38, 6689–6695 (2011)
[9] Negri, P.: Estimating the queue length at street intersections by using a movement feature space approach. IET IP 8(7), 406–416 (2014)
[10] Negri, P., Goussies, N., Lotito, P.: Detecting pedestrians on a movement feature space. PR 47(1), 56–71 (2014)
[11] Rosales-Pérez, A.e.a.: Surrogate-assisted multi-objective model selection for support vector machines. Neurocomputing 150, 163–172 (2015)
[12] Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR. vol. 1, pp. 511–518 (2001)