Multiple Instance Boosting for Face Recognition in Videos

Paul Wohlhart, Martin Köstinger, Peter M. Roth, and Horst Bischof

Institute for Computer Graphics and Vision
Graz University of Technology, Austria
{*wohlhart,koestinger,pmroth,bischof*}*@icg.tugraz.at*

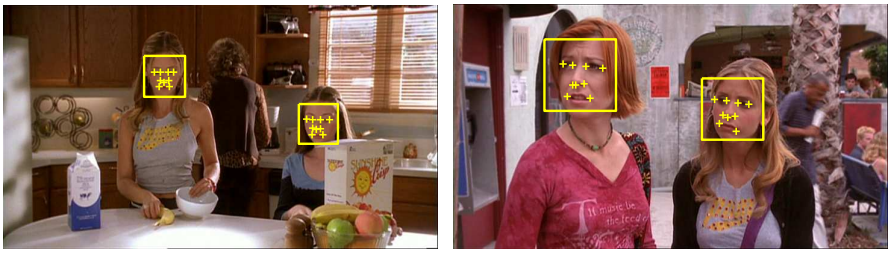
Abstract. For face recognition from video streams often cues such as transcripts, subtitles or on-screen text are available. This information could be very valuable for improving the recognition performance. However, frequently this data can not be associated directly with just one of the visible faces. To overcome this limitations and to exploit valuable information, we define the task as a multiple instance learning (MIL) problem. We formulate a robust loss function that describes our problem and incorporates ambiguous and unreliable information sources and optimize it using Gradient Boosting. A new definition of the posterior probability of a bag, based on the L_p-norm, improves the ability to deal with varying bag sizes over existing formulations. The benefits of the approach are demonstrated for face recognition in videos on a publicly available benchmark dataset. In fact, we show that exploring new information sources can drastically improve the classification results. Additionally, we show its competitive performance on standard machine learning datasets.

1 Introduction

TV and video-sharing websites constantly provide large amounts of digital video data. This data could be an extremely valuable and important source of information, that today remains mostly unexplored. In fact, since most of the video data is only indexed by some meta-data and not by its content, it is inaccessible to goal-oriented search. Manual annotation is laborious or even infeasible at large scale, thus, to allow for a more efficient search and retrieval, methods for automatic interpretation of the visual content are needed.

In this paper, we address the problem of fully automated identification of people in videos, where we have to carry out the following steps: First, detecting people's faces and tracking them throughout a scene. Second, automatically extracting as much information as possible about the persons' identities from associated information sources, such as the audio track (speech recognition), subtitles, the transcript, on-screen text, or electronic program guide (EPG) data. Third, using the gathered data to learn to re-identify them in different contexts, only based on their visual appearance.

This problem was recently tackled by several authors [3, 4, 7, 8, 17, 18]. Everingham *et al.* [7, 8] label exemplars by visual speaker detection. The name of



(a) Buffy, Dawn, Joice (b) Willow, Buffy

Fig. 1. Face recognition in videos: Often valuable information cannot be assigned unambiguously to exactly one person. For instance we know from the video transcript that a character is present in a scene, but the corresponding face is unknown.

the speaker is obtained by automatically aligning the timing information of the subtitles with the naming information from the transcript. However, due to the nearest neighbor classification label noise is propagated. Thus, the method cannot recover from labeling errors. The work of Sivic *et al.* [18] replaces the nearest neighbor framework by multiple kernel classification. The base kernels operate on the min-min distance between HOG blocks. Therefore, the optimized combination coefficients describe the relative importance of the individual blocks for classification. Nevertheless, it is not possible to integrate cues providing information that can not be assigned unambiguously to one single instance. Ramanan *et al.* [17] use a multitude of inference cues to obtain face clusters. Different cues apply to different time scales. However, the system requires manual user interaction to label an initial set of face clusters.

Thus, these methods require either manual labeling or cannot make use of information that applies to multiple instances. However, this is a reasonable scenario when learning from videos and associated sources. For instance, as illustrated in Figure 1, we know from textual cues that a specific character should be present in one scene of a movie. But we do not know to which of the currently visible faces this information corresponds. The goal of this paper is to make use of information which cannot be disambiguated. Additionally, we have to ensure robustness, *i.e.*, since the information extraction procedure is not completely reliable, we have to inherently deal with noisy and uncertain labels. We meet these requirements by formulating the task as a Multiple Instance Learning (MIL) problem.

In particular, for that purpose we adopt Gradient Boosting. Compared to other methods Gradient Boosting has the advantage that any loss function that fits the task can be used, as long as it is differentiable, thus providing a very general optimization framework. In our case we build on the Logit-loss function – to ensure the required robustness – and further incorporate the MIL constraints. The approach is similar to the one of Viola *et al.* [19], however, their formulation implicitly assumes that all bags in the training data are more or less of the same

size and essentially not too big. To overcome this limitation, we define a new formulation of the posterior probability of a bag, approximating more directly the original definition of MIL, which is better suited for our task. Additionally, we generalize the framework such that arbitrary learning algorithms can be used to form the weak hypotheses.

In the following, we first introduce our new Gradient Boosting based MIL algorithm and then give an experimental evaluation on both, standard benchmark datasets as well as on a publicly available face recognition dataset.

2 MIL - Boosting

In a supervised learning scenario the training data is given in the form of a set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is a sample and $y_i \in \mathcal{Y} = \{-1, +1\}$ its corresponding binary label. However, in practice, it is often hard or even impossible to assign a label to all samples. But it is rather easy to specify a group of data samples for which it can be ensured that at least one instance carries the label, which leads to Multiple Instance Learning (MIL) [6]. In MIL the data is provided in form of labeled bags $\mathcal{D}_{\text{mil}} = \left\{(\mathcal{B}_1^l, y_1), \ldots, (\mathcal{B}_N^l, y_N)\right\}$, where $\mathcal{B}_i = \{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iN_{B_i}}\}$, $\mathbf{x}_{ij} \in \mathbb{R}^d$, is a bag containing N_{B_i} samples and $y_i \in \mathcal{Y}$ its binary label. A bag is defined to be positive if *at least one* instance in the bag is positive, whereas accordingly for a negative bag all instances have to be negative. Building on these ideas, in the following, we will derive a new formulation for MIL which is based on Gradient Boosting.

2.1 Gradient Boosting

In general, the goal of Boosting is to estimate a strong classifier $F(\mathbf{x})$ as a linear combination of weak classifiers $f_t(\mathbf{x})$ such that the the expected classification error is minimized:

$$F(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t f_t(\mathbf{x}) \ . \tag{1}$$

In particular, Gradient Boosting aims to find a strong classifier $F^*(\mathbf{x})$ by solving the following optimization problem:

$$F^*(\mathbf{x}) = \arg\min_{F(\mathbf{x})} \mathcal{L}(\mathcal{D}; F(\mathbf{x})) \ , \tag{2}$$

where $\mathcal{L}(\mathcal{D}; F(\mathbf{x}))$ is a *loss function* measuring the performance of the classifier by giving penalties for misclassified training examples.

Gradient Boosting iteratively estimates the function $F^*(\mathbf{x})$ by greedily constructing base functions $f_t(\mathbf{x})$ (weak learners) based on the preceding $f_1(\mathbf{x}), \ldots, f_{t-1}(\mathbf{x})$. This is accomplished by taking the derivative of the loss function with respect to the current strong classifier's output for each training sample and constructing the new $f_t(\mathbf{x})$ such as to produce outputs that approximate the inverse direction of this gradient (*i.e.*, reduce the residuals):

$$f_t(\mathbf{x}) = \arg\max_{f(\mathbf{x})} \left\langle -\left\{ \frac{\partial \mathcal{L}(\mathcal{D}; F)}{\partial F(\mathbf{x}_1)}, \ldots, \frac{\partial \mathcal{L}(\mathcal{D}; F)}{\partial F(\mathbf{x}_N)} \right\}, \{f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)\} \right\rangle \quad (3)$$

$$= \arg\max_{f(\mathbf{x})} -\sum_{i=1}^{N} \frac{\partial \mathcal{L}(\mathcal{D}; F)}{\partial F(\mathbf{x}_i)} f(\mathbf{x}_i) \ . \quad (4)$$

Finally, when the new $f_t(\mathbf{x})$ is found, the best weight α_t is determined by a line search.

2.2 Loss Functions

The main advantage of Gradient Boosting over other Boosting variants is the flexibility of choosing a loss function that suites the task to be solved. Several different losses have been proposed in the literature (Exponential [10], Logit [11], Savage [16]), mainly differing in the way how misclassified samples are punished, mainly influencing the robustness of the method against label noise. Since the Logit loss has shown to be a considerable trade-off between robustness and performance we build our algorithm on it. Thus, in the following we derive a Gradient Boosting variant using a Logit loss, which can then easily be extended by incorporating the Multiple Instance Learning constraints in Section 2.3.

The Logit loss of a classifier $F(\mathbf{x})$ over a dataset $\mathcal{D}$ is defined as

$$\mathcal{L}(\mathcal{D}; F(\mathbf{x})) = \sum_{i=1}^{N} \log\left(1 + e^{-y_i F(\mathbf{x}_i)}\right) = -\sum_{i=1}^{N} \log\left(\frac{1}{1 + e^{-y_i F(\mathbf{x}_i)}}\right) \ . \quad (5)$$

Thus, taking the logistic regression of the strong classifier's output $F(\mathbf{x})$, let

$$P(y{=}z|\mathbf{x}_i) = \frac{1}{1 + e^{-zF(\mathbf{x}_i)}} \quad (6)$$

be the predicted probability that an instance $\mathbf{x}$ is assigned the label $z \in \mathcal{Y}$. Then, we can interpret Eq. (5) as the cross entropy of the labels and the instance probabilities reported by the classifier:

$$\mathcal{L}(\mathcal{D}; F(\mathbf{x})) = -\sum_{i=1}^{N} \sum_{z \in \mathcal{Y}} [z{=}y_i] \log\left(P(y{=}z|\mathbf{x}_i)\right) \ , \quad (7)$$

where $[\cdot]$ is the Iverson bracket.

With this loss the optimization for the weak learners in Eq. (4) becomes

$$f_t(\mathbf{x}) = \arg\max_{f(\mathbf{x})} \sum_{i=1}^{N} \sum_{z \in \mathcal{Y}} [z{=}y_i] \frac{\partial \log P(y{=}z|\mathbf{x}_i)}{\partial F(\mathbf{x}_i)} f(\mathbf{x}_i) \ . \quad (8)$$

Thus, we are looking for a new $f_t(\mathbf{x})$ whose output approximates the derivative of the log of the instance probabilities, which we denote as

$$a_i(z) = \frac{\partial \log P(y=z|\mathbf{x}_i)}{\partial F(\mathbf{x}_i)} \ . \tag{9}$$

Generally, existing learning algorithms are not designed to solve Eq. (8). However, we can define a weight w_i for each training sample as

$$\forall \mathbf{x}_i \in \mathcal{D}: \quad w_i = |a_i(y_i)| \ . \tag{10}$$

Thus, we are very flexible and can use any learning algorithm that can handle training data with (importance-)weighted samples to construct a new weak learner approximating the gradient.

2.3 Solving MIL with Gradient Boosting

In order to solve the MIL problem we define a new loss function over the bags

$$\mathcal{L}(\mathcal{D}_{\text{mil}}; F(\mathbf{x})) = -\sum_{i=1}^{N} \sum_{z \in \mathcal{Y}} [z=y_i] \log(P(y=z|\mathcal{B}_i)) \ , \tag{11}$$

where $P(y=1|\mathcal{B}_i)$ is the bag posterior. Following the definition of MIL, the bag posterior is defined over the probabilities of its instances as

$$P(y=1|\mathcal{B}_i) = \max_j P(y=1|\mathbf{x}_{ij}) \ . \tag{12}$$

However, this measure is not differentiable, thus, approximations have to be used. For instance, Viola *et al.* [19] proposed to use *noisy-or* [15] as the bag posterior model:

$$P_{\text{NOR}}(y=1|\mathcal{B}_i) = 1 - \prod_{j=1}^{N_{B_i}} (1 - P(y=1|\mathbf{x}_{ij})) \ . \tag{13}$$

The main disadvantage of the *noisy-or* formulation is that the size of the bag (number of instances) substantially influences the outcome. For example, if all instances in a bag have a very low probability, it is still assigned a high posterior probability if the number of instances is large. This is especially unfavorable if the size of the bags varies strongly within the training data, as it is the case in our task.

Therefore, we propose to use a more direct approximation to the max operation in Eq. (12), by making use of the L_p-norm:

$$P_{L_p}(y=1|\mathcal{B}_i) = \left(\sum_{j=1}^{N_{\mathcal{B}_i}} P(y=1|\mathbf{x}_{ij})^p \right)^{1/p} \ . \tag{14}$$

For large values of p this well approximates the max operation and as $p \to \infty$ even converges to it. Thus, according to Eq. (8), the optimization for generating the next weak learner is given by

$$f_t(\mathbf{x}) = \arg\max_{f(\mathbf{x})} \sum_{i=1}^{N} \sum_{z \in \mathcal{Y}} [z = y_i] \sum_{j=1}^{N_{\mathcal{B}_i}} \frac{\partial \log P(y = z|\mathcal{B}_i)}{\partial F(\mathbf{x}_{ij})} f(\mathbf{x}_{ij}) \ . \tag{15}$$

Again, we can derive the weights for each instance by

$$\forall (\mathcal{B}_i, y_i) \in \mathcal{D}_{\text{mil}}, \forall \mathbf{x}_{ij} \in \mathcal{B}_i : \quad w_{ij} = |a_{ij}(y_i)| \ , \tag{16}$$

where, in contrast to Eq. (9), the $a_{ij}(z)$ are now defined on bag level:

$$a_{ij}(z) = \frac{\partial \log P(y = z|\mathcal{B}_i)}{\partial F(\mathbf{x}_{ij})} \ . \tag{17}$$

In our case, the derivation of Eq. (14) is given by

$$a_{ij}^{L_p}(z) = \frac{\hat{z} - P(y=1|\mathcal{B}_i)}{1 - P(y=1|\mathcal{B}_i)} (1 - P(y=1|\mathbf{x}_{ij})) \frac{P(y=1|\mathbf{x}_{ij})^p}{\sum_{k=1}^{N_{\mathcal{B}_i}} P(y=1|\mathbf{x}_{ik})^p} \ , \tag{18}$$

where $\hat{z} = (z+1)/2$. The bigger we choose p the better the approximation. As $p \to \infty$, we get

$$\tilde{a}_{ij}^{L_\infty}(z) = (\hat{z} - P(y=1|\mathcal{B}_i)) \left[P(y=1|\mathbf{x}_{ij}) = \max_k P(y=1|\mathbf{x}_{ik}) \right] / N_{\mathcal{B}_i,\max} \ , \tag{19}$$

where $N_{\mathcal{B}_i,\max} = |\{j | P(y=1|\mathbf{x}_{ij}) = \max_k P(y=1|\mathbf{x}_{ik})\}|$ is the number of instances in bag $\mathcal{B}_i$ having the highest probability. Note that $\tilde{a}_{ij}^{L_\infty}$ is not necessarily the analytical derivative of P_{L_∞}, since the series of P_{L_p} converges pointwise, but not uniform. Nevertheless, we use it since it gives the best approximation for the weights w_{ij} and it is easy to compute.

3 Benchmark Datasets

Before showing results for the actual task, i.e., face recognition, we would like to give a broad quantitative comparison to other methods. In particular, we evaluate the proposed MILBoost using the P_{L_p} bag posterior model on the well known and frequently used CBIR machine learning database [1] with its three multiple instance datasets *Tiger*, *Fox* and *Elephant* as well as on the two Musk datasets [6]. Here, as well as in the other experiments, the weak learners used are probabilistic decision stumps, which test one feature of a sample against a threshold and report a probability of begin positive, estimated from the training data, on either side. The mean areas under the ROC curves over 10 individual 10-fold cross validation runs are reported in Table 1.[1]

[1] Note that for mi-SVM and MI-SVM there are three different versions depending on the kernel (linear, poly, rbf) and we report the best one for each class.

On Musk, MILBoost is in the range of state-of-the-art algorithms, although it does not reach the performance of certain specialized methods. However, MILBoost with the noisy-or bag posterior model, to the best of our knowledge, delivers the best results reported so far for the Tiger and Elephant classes of the CBIR dataset. Our P_{L_p} bag posterior model also produces very good results on those two classes and considerably outperforms noisy-or on the difficult Fox dataset. Note also, that its theoretical advantage of being able to handle variably sized bags does not apply for these datasets, since the bags are of equal size.

Table 1. Results of various MIL algorithms on the standard MIL datasets CBIR and Musk1&2. MILBoost outperforms all other methods on CBIR, with MILBoost L_p producing the best overall performance. The best performance for each dataset is marked in bold, second best in italics.

	Tiger	**Elephant**	**Fox**	**Musk1**	**Musk2**
sbMIL [2]	82.95	88.58	*69.78*	*91.78*	87.40
NSK [13]	79.07	82.94	64.01	85.61	**90.78**
MI-SVM [1]	84.00	81.40	59.40	81.50	86.30
mi-SVM [1]	78.90	82.20	58.20	87.40	83.60
MI-CRF [5]	78.90	82.20	58.20	77.90	84.30
PPMM [20]	80.20	82.00	60.30	**95.60**	81.20
MICA [12]	82.00	82.50	62.00	84.40	*90.50*
ALP-SVM [14]	86.00	83.50	66.00	86.30	86.20
MILBoost n-or	**91.70**	**93.43**	65.72	81.98	81.92
MILBoost L_p	*89.79*	*91.82*	**71.80**	81.98	81.87

4 Face Recognition from Videos

In the following we demonstrate our method for face recognition from associated information sources on the publicly available part of the *Buffy* dataset proposed by Everingham *et al.* [7][2]. It consists of 27504 individual frontal face detections and additionally provides face descriptions and face tracks. Faces are described by normalized pixel patches extracted at salient facial feature points, which are localized by a Pictorial Structures model [9]. Within a shot face detections (in individual frames) are grouped into face tracks by motion information. Hence, the task is to assign the correct cast name to each of the 516 face tracks. The cast list of the ground truth annotation consists of 11 named entities, the class *other* and *false positive* of the detection process. For each cast member we train a one-vs.-all classifier.

To automatically obtain training labels, we exploit information sources closely associated to the video, namely transcript and subtitles, both containing the dialogs. The transcript additionally provides naming information and embraces scenes with a textual description of what is happening. From the transcript we

[2] The more recent "Buffy" dataset [18] is not publicly available.

extract the coarse scene structure. Further, to augment the transcript with the timing information it is aligned with the subtitles by dynamic time warping. Thus, we now know who is speaking when but neither if the speaker is visible or to which face the current utterance belongs.

We use these cues to compose training bags. A bag consists of one or more face tracks and an associated label. First, we form *speaker bags*. To judge if a person is speaking we observe the optical flow [21] around the mouth region. Tracks identified as speaking are assigned the label of the current speaker from the augmented transcript. Second, we define *scene bags* that contain all face tracks present in a scene. The idea is to decide if a certain character is likely to appear in a particular scene or not, dependent on the number of spoken text chunks. To finally test the labeling performance, each face track forms a singleton bag. Testing is done standalone based on pure face appearance and does not need additional information.

Compliant with previous work we measure the performance in a *refusal to predict* style. By taking the difference of the leading two classifier scores a confidence is obtained. Further, we rank and threshold the confidences. In that sense, recall means the percentage of face tracks which have a higher confidence than the current threshold and thus are labeled. Precision means the ratio of correctly labeled samples. We first report the performances of the different models for the bag posterior probabilities on this task. The comparison is shown in Table 2, where it can be seen that, as expected, P_{L_p} outperforms P_{NOR} over most levels of recall, especially for higher recall values. Thus, for the succeeding experiments we just use the P_{L_p} bag posterior.

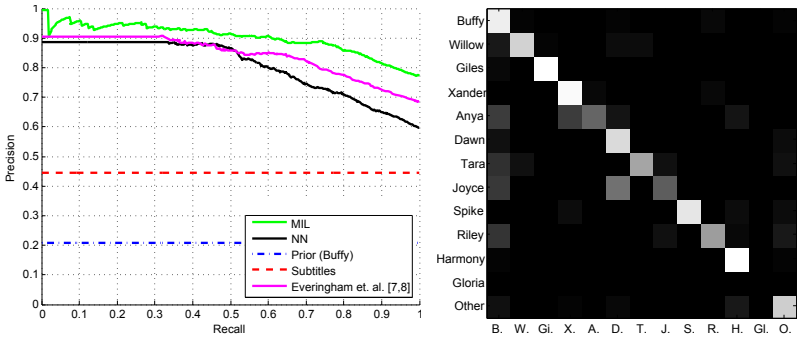
Table 2. Performance comparison of the different models for the posterior probability of a bag. P_{L_p} outperforms P_{NOR} over most levels of recall.

Recall	50%	60%	70%	80%	90%	100%
P_{L_p}	91,5%	**90,9%**	**88,7%**	**86,3%**	**81,8%**	**77,7%**
P_{NOR}	91,5%	90,6%	86,5%	83,9%	78,5%	73,8%

Next, in Figure 2 we compare our method with previous work [7, 8]. Everingham *et al.* proposed to classify each track based on the min-min distance to the tracks labeled by the speaker detection. The min-min distance $d_f(F_i, F_j)$ between two face tracks F_i and F_j is defined as:

$$d_f(F_i, F_j) = \min_{f_i \in F_i} \min_{f_j \in F_j} \|f_i - f_j\| , \tag{20}$$

where $f_i \in F_i$ and $f_j \in F_j$ are face descriptions. This method is denoted as NN. For comparison, we also include the original curve from [7]. Please note that this method makes use of additional clothing descriptors and a different speaker detection, not provided with the published dataset. As reference we also state the performance of labeling all face tracks with the cast name appearing most frequently in the transcript (Prior on *Buffy*). Further, also the performance of using the aligned subtitles to propose a name is reported. With the speaker



(a) Precision/Recall (b) Confusion Matrix

Fig. 2. Buffy dataset: (a) MIL-L_p clearly outperforms the baseline (NN) over all levels of recall. Subtitles describe further baseline methods, see text for details. (b) The associated confusion matrix.

detection we can label 33.4% of the tracks with a precison of 89.0%. Please note that the baseline method provides no means for ranking the tracks detected as *speaking*. Therefore, the curve is constant for the first levels of recall. Due to the nearest neighbor classification the baseline method has no real chance to recover from labeling errors. Label noise propagates directly into the classification. If the method is required to label all face tracks a precision of 60.1% is reached. MIL clearly outperforms the baseline method over all levels of recall. At 100% recall the precision is 77.7%. This is an improvement of 17.6% over the baseline. Indeed, the method even delivers a higher precision than the speaker detection up to a recall level of 65%. It labels nearly twice as many face tracks with an accuracy of 89%. This shows clearly the ability of MIL to recover from labeling errors.

5 Conclusion

In this work we presented the task of face recognition in weakly labeled videos as Multiple Instance Learning problem. We formulated the MIL concept in a probabilistic loss function and optimized it in a Gradient Boosting framework. The new formulation of the posterior probabilities of the bags using the L_p-norm allows us to better deal with bags of varying size, as the comparison with noisy-or confirmed. The evaluation on standard machine learning data shows excellent results for the learning algorithm. Further, the task of face recognition in videos verified that it is able to benefit from ambiguous and even noisy data. This can be attributed to the design of the loss function, based on Logit. It gives penalties for misclassifying training samples, but does not exaggerate the influence of very wrong classifications to avoid over-fitting to potentially noisy labels.

Acknowledgments. The work was supported by the FFG projects MDL (818800) and SECRET (821690) under the Austrian Security Research Programme KIRAS.

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Advances NIPS (2003)
2. Bunescu, R.C., Mooney, R.J.: Multiple instance learning for sparse positive bags. In: Proc. ICML (2007)
3. Cour, T., Sapp, B., Jordan, C., Taskar, B.: Learning from ambiguously labeled images. In: Proc. CVPR (2009)
4. Cour, T., Sapp, B., Nagle, A., Taskar, B.: Talking pictures: Temporal grouping and dialog-supervised person recognition. In: Proc. CVPR (2010)
5. Deselaers, T., Ferrari, V.: A conditional random field for multiple-instance learning. In: Proc. ICML (2010)
6. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence 89(1–2), 31–71 (1997)
7. Everingham, M., Sivic, J., Zisserman, A.: "Hello! My name is... Buffy" – automatic naming of characters in TV video. In: Proc. BMVC (2006)
8. Everingham, M., Sivic, J., Zisserman, A.: Taking the bite out of automatic naming of characters in TV video. Image and Vision Computing 27(5) (2009)
9. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. Intern. Journal of Computer Vision 61, 55–79 (2005)
10. Freund, Y., Shapire, R.E.: Experiments with a new boosting algorithm. In: Proc. ICML (1996)
11. Friedman, J., Hastie, T., Tibshirani., R.: Additive logistic regression: a statistical view of boosting. The Annals of Statistics 28(2), 337–374 (2000)
12. Fung, G., Rosales, R., Krishnapuram, B.: Learning rankings via convex hull separation. In: Advances NIPS (2006)
13. Gärtner, T., Flach, P.A., Kowalczyk, A., Smola, A.J.: Multi–instance kernels. In: Proc. ICML (2002)
14. Gehler, P.V., Chapelle, O.: Deterministic annealing for multiple-instance learning. In: Proc. Int. Conf. on Artificial Intelligence and Statistics (2007)
15. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. In: Advances NIPS (1998)
16. Masnadi-Shirazi, H., Vasconcelos, N.: On the design of loss functions for classification: theory, robustness to outliers, and SavageBoost. In: Advances NIPS (2009)
17. Ramanan, D., Baker, S., Kakade, S.: Leveraging archival video for building face datasets. In: Proc. ICCV (2007)
18. Sivic, J., Everingham, M., Zisserman, A.: "Who are you?" – learning person specific classifiers from video. In: Proc. CVPR (2009)
19. Viola, P., Platt, J., Zhang, C.: Multiple instance boosting for object detection. In: Advances NIPS (2006)
20. Wang, H.Y., Yang, Q., Zha, H.: Adaptive p-posterior mixture-model kernels for multiple instance learning. In: Proc. ICML (2008)
21. Werlberger, M., Trobin, W., Pock, T., Wedel, A., Cremers, D., Bischof, H.: Anisotropic Huber-L1 optical flow. In: Proc. BMVC (2009)